

Cortical free association dynamics: distinct phases of a latching network

Eleonora Russo¹ and Alessandro Treves¹

¹SISSA, Cognitive Neuroscience, via Bonomea 265, 34136 Trieste, Italy*

(Dated: February 25, 2013)

A Potts associative memory network has been proposed as a simplified model of macroscopic cortical dynamics, in which each Potts unit stands for a patch of cortex, which can be activated in one of S local attractor states. The internal neuronal dynamics of the patch is not described by the model, rather it is subsumed into an effective description in terms of graded Potts units, with adaptation effects both specific to each attractor state and generic to the patch. If each unit, or patch, receives effective (tensor) connections from C other units, the network has been shown to be able to store a large number p of global patterns, or network attractors, each with a fraction a of the units active, where the critical load p_c scales roughly like $p_c \approx CS^2/a \ln(1/a)$ (if the patterns are randomly correlated). Interestingly, after retrieving an externally cued attractor, the network can continue jumping, or *latching*, from attractor to attractor, driven by adaptation effects. The occurrence and duration of latching dynamics is found through simulations to depend critically on the strength of local attractor states, expressed in the Potts model by a parameter w . Here we describe with simulations and then analytically the boundaries between distinct phases of no latching, of transient and sustained latching, deriving a phase diagram in the plane $w - T$, where T parametrizes thermal noise effects. Implications for real cortical dynamics are briefly reviewed in the conclusions.

I. INTRODUCTION

The Hopfield [1] model and its subsequent analysis with statistical physics techniques [2] have provided an important quantitative theoretical framework to approach the operation of associative memory networks in the brain. After many years of modeling developments and gradual familiarization by the neuroscience community [3, 4], the notion of attractor states forcefully put forward by the Hopfield model has guided experimental investigations in the rat hippocampus [5, 6] and in the primate inferotemporal cortex [7, 8], suggesting analyses of the time course of attractor dynamics which have revealed a striking rhythmically pulsed or "flickering" regime [9].

Attractor dynamics has been proposed to be relevant to global cortical function [10, 11], also in situations in which the cortex actually produces behavior, e.g. speech, confabulation, or a drawing, as opposed to just responding to a stimulus from outside [12–18]. It appears then of particular interest to go, within a modeling framework, beyond the standard cued retrieval operation, to protracted cortical dynamics including attraction to, and escape from, a sequence of metastable states, or attractor "ruins" [19, 20]. Firing frequency adaptation is a simple and salient feature of neuronal dynamics in the cortex, which results in the unavoidable metastability of any of its attractor states. The study of a full neuronal network model of an extended "cortex" with firing frequency adaptation is complicated, however, by the necessity to treat simultaneously local and global dynam-

ics [21]. The Potts associative memory network model [22, 23] can be interpreted as a global cortical model, in which local dynamics are subsumed into an effective description [24, 25]. Its storage and retrieval capacity (without adaptation) can be studied analytically [26]. With adaptation, it shows latching behavior, i.e. a sequence of jumps from one attractor to another (Fig. 1), with non-trivial transition statistics, neither random nor deterministic [27, 28]. The aim of the present study is

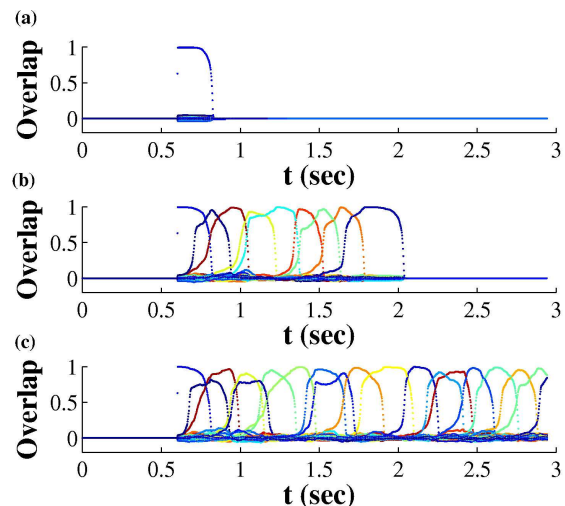


Figure 1: (Color online) Examples of latching dynamics in the Potts network. The correlation (overlap) of the network state with each of a number of stored activity patterns is indicated in a different color. The retrieval of a memory can be followed by the inactive state (a) or by latching dynamics, that either die out after a few steps (b) or self-sustain indefinitely (c), depending on the parameters of the network.

*URL: <http://www.sissa.it/~ale/limbo.html>; Electronic address: russo@sisssa.it

to characterize the conditions for latching dynamics to occur, and to self-sustain indefinitely. The model is defined in the next section, and the various latching regimes are described in Sec. III. Then in Sec. IV an analytical approach is developed, that yields a phase diagram for the model. Its implications for real cortical dynamics are reviewed in the last section.

II. THE MODEL

The network considered here is the same as studied previously [24] and we refer to that paper and to Ref. [25] for a discussion of the correspondence with real cortical networks. Here we dwell on the mathematical description of the model, based on the hypothesis that, abstracted from observed features of cortical representations [29], it may be relevant to a better understanding of cortical dynamics, over time scales of up to a few seconds.

A. Potts units

Let us consider a network comprised of many subnetworks interacting with each other with long range connections. Each of these small networks has its own attractors, with which it tends spontaneously to align during the dynamics. So, while in a brief transient condition the activation state of each small network may be partially overlapping with several of its attractors, during a protracted quasi-stable memory retrieval condition its alignment with one of the attractors tends to be near complete.

To summarily describe these local dynamics we represent the small subnetworks with Potts units, each of which can be active, to a variable degree from 0 to 1, in a number S of distinct active states, representing local attractors, and in one inactive state, when the local activity pattern has no correlation with any of the attractors (Fig. 2). At any time, the sum of the activity levels of unit i , σ_i^k , $k \in [0, S]$, ($k = 0$ denotes the inactive state) is kept fixed to 1, $\sum_{k=0}^S \sigma_i^k = 1$.

The model thus reduces to an autoassociative network of N Potts units bound by long range connections. The connections, stored in the weight matrix of the network, are fixed, with the standard assumption [1] that they reflect an earlier learning phase. By setting the weights, one defines the global attractors of the system. These are configurations of states of Potts units, that represent combinations of local attractors in each subnetwork, and in turn serve as attractors for the global dynamics of the Potts units. In particular we take the network to have stored as global attractors p activity patterns ξ^μ (see Sec. IID), with $\mu = 1, \dots, p$. The strength J_{ij}^{kl} of the connection between state k of unit i and state l of

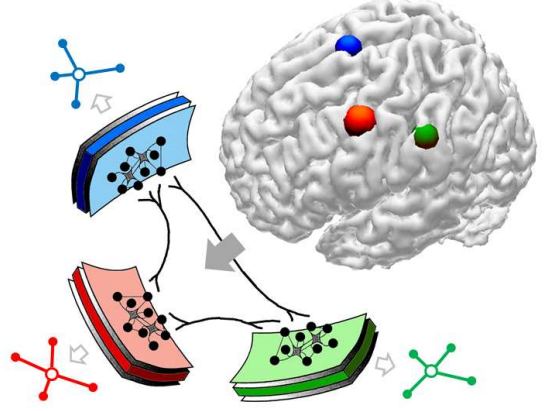


Figure 2: (Color online) Conceptual derivation of the Potts network, obtained as a simplification of a two-tier associative memory network by reducing its local dynamics to single Potts units, representative of the states of patches of cortex.

unit j reads [23, 26]

$$J_{ij}^{kl} = \frac{c_{ij}}{Ca(1 - \frac{a}{S})} \sum_{\mu=1}^p (\delta_{\xi_i^\mu k} - \frac{a}{S})(\delta_{\xi_j^\mu l} - \frac{a}{S})(1 - \delta_{k0})(1 - \delta_{l0}) \quad (1)$$

where C is the average number of units connected with a given unit (denoted as c_M in [24] to indicate connectivity between Modules) and a is the *sparsity* parameter, i.e. the fraction of units in an active state in each of the stored patterns [24]. The connection matrix element c_{ij} is equal to 1 if the presynaptic unit (module) j is connected to the postsynaptic unit (module) i , and 0 otherwise. The connection tensor J_{ij}^{kl} can be written more succinctly by dropping the last two factors and intending it to act only between active states, given that the inactive state of each unit, σ_i^0 , is not an independent variable because of the constraint $\sum_{k=0}^S \sigma_i^k = 1$.

A simple update rule can be assigned to the network in discrete time, by randomly selecting each unit at a time, evaluating the *local field* in the direction of its different states $h_i^k = \sum_j J_{ij}^{kl} \sigma_j^l$ (plus any input from outside the Potts network), and aligning the unit in the direction k^* of the strongest field if this surpasses a threshold U , $\sigma^{k^*} = 1$, and setting it in the inactive state, $\sigma^0 = 1$, otherwise. The thermodynamics of such a model, for $c_{ij} \equiv c_{ji}$, has been analysed by [26], deriving it from the Hamiltonian

$$H = -\frac{1}{2} \sum_{i,j \neq i}^N \sum_{k,l=1}^S J_{ij}^{kl} \sigma_i^k \sigma_j^l + U \sum_i^N \sum_{k=1}^S \sigma_i^k \quad (2)$$

which can be thought to act under the influence of noise, modeled by introducing a non-zero temperature $T \equiv \beta^{-1}$. The network stabilizes into one of several possible asymptotic states, the global attractors, which may include clusters of configurations around the quiescent

state $\{\sigma_i^0 \equiv 1\}$ as well as around the memory attractors $\{\sigma_i^{\xi^\mu} = 1\}$ (all other states of a unit in the low noise regime being at $\sigma_i \simeq 0$). For $T = \beta^{-1} = 0$ (no thermal noise at all) the asymptotic states reduce to a single configuration, and no further updating occurs.

We aim here for a more realistic description of local cortical dynamics, which occurs in continuous time and includes a role for neuronal fatigue, to see the resulting effects on global dynamics.

B. Adaptive dynamics

Transitions among attractors may be induced by a variety of mechanisms that either weaken temporarily the strength of the current attractor or produce an enhanced cue in the direction of another one [30–37].

The stability of the stored memories is weakened, in the model considered here, by the adaptive thresholds introduced to model neuronal fatigue and slow inhibition. Two types of time-varying thresholds are included: an additional contribution to the constant baseline threshold U , affecting all active states of a Potts unit, $\theta_i^0(t)$, with a time constant τ_3 of hundreds of ms for its dynamics, intended to model various forms of slow, delayed inhibition within a cortical patch [38–40], and a specific threshold $\theta_i^k(t)$, affecting only Potts state k , with a time constant τ_2 of tens of ms for its dynamics, intended to model resource depletion of the neurons (fatigue proper) and of the synapses (short-term depression [41]) active in that state. With these modifications, after the stabilization into a global pattern, all the active Potts units are increasingly affected by fatigue and, with two characteristic time constants, tend to inactivate or to change active state, eventually changing the activity configuration of the network. Thereby the system is pushed away from the basin of attraction it had settled in and, sometimes, towards that of a new correlated memory attractor.

The activity of each unit is determined by a set of specific input variables $r_i^k(t)$, which rapidly integrate, with time constant τ_1 (meant to be in the ms range), the *local fields*, i.e. the summed influence $h_i^k(t)$ of presynaptic units, subject to the dynamic thresholds $\theta_i^k(t)$, while the non-specific thresholds $\theta_i^0(t)$ and U can be thought of as inputs to the inactive state. In detail, the activation of the unit i is set at any time t by the non-linearity

$$\sigma_i^k = \frac{\exp(\beta r_i^k)}{\sum_{l=1}^S \exp(\beta r_i^l) + \exp[\beta(\theta_i^0 + U)]} \quad (3)$$

for each active state k and

$$\sigma_i^0 = \frac{\exp[\beta(\theta_i^0 + U)]}{\sum_{l=1}^S \exp(\beta r_i^l) + \exp[\beta(\theta_i^0 + U)]} \quad (4)$$

for the null state, with β an inverse temperature parametrizing thermal noise, as in the Hopfield model [1]. Note that $T \equiv \beta^{-1} > 0$ implies a smooth transfer function for the Potts units, but no stochasticity in their dynamics. The temperature in other words represents here

thermal noise in the local network underlying each Potts unit, but the global network is noiseless. See [42] for the analysis of a related model but with different, noisy dynamics.

The inputs are linearly integrated

$$\tau_1 \frac{dr_i^k(t)}{dt} = h_i^k(t) - \theta_i^k(t) - r_i^k(t). \quad (5)$$

with

$$h_i^k = \sum_{j \neq i} \sum_{l=1}^S J_{ij}^{kl} \sigma_j^l + w(\sigma_i^k - \frac{1}{S} \sum_{l=1}^S \sigma_i^l) \quad (6)$$

where it is important to note that we have added a self-reinforcement term with coefficient w , to model the non-linear convergence towards the more active state, induced by the local patch dynamics not explicitly represented in the Potts model. The w term is then a key difference between the original Potts autoassociative network model, as studied by Kanter [22], and our version, intended as an effective model of a two-tier network. The parameter w provides a positive feedback, that makes the convergence towards a local state a rapid self-regenerative process; it can be argued to numerically reflect the ratio of local to external afferent inputs, to each neuron in the two-tier network [43]. We shall call it the *local feedback* term.

The adaptive thresholds follow, with slower dynamics, the mean activity of the unit

$$\tau_2 \frac{d\theta_i^k(t)}{dt} = \sigma_i^k(t) - \theta_i^k(t) \quad (7)$$

and

$$\tau_3 \frac{d\theta_i^0(t)}{dt} = \sum_{k=1}^S \sigma_i^k(t) - \theta_i^0(t). \quad (8)$$

These equations completely define the dynamics of the network.

Note that in the limit $\tau_2, \tau_3 \rightarrow \infty$ the model sheds the adaptive character of its dynamics, and genuine attractor states are indefinitely stable.

Stability is also maintained when attractors are exceedingly "deep", that is, if the local feedback w is strong enough to overcome the effects of the rising thresholds. Then the thresholds stabilize at the asymptotic values $\theta_i^k \rightarrow \bar{\sigma}_i^k$ and $\theta_i^0 \rightarrow \sum_k \bar{\sigma}_i^k$, and the input variables $\{r\}$'s satisfy the asymptotic system $r_i^k = h_i^k - \theta_i^k$, where both the $\{h\}$'s and $\{\theta\}$'s are functions of the $\{\bar{\sigma}\}$'s, hence of the $\{r\}$'s through Eqs. 3. Away from those limit cases, adaptation tends to destabilize memory attractors.

The one state that is guaranteed to remain stable (provided the constant threshold U takes a positive value) is the "global null state", where all active states take low values, vanishing exponentially with β , and $\sigma_i^0 \simeq 1$ (the exact values depend on β and U).

C. The slowly adaptive regime

The presence of adaptive thresholds makes an analytical approach to the study of the dynamics quite difficult, in general. Were it not for the adaptive thresholds, and assuming symmetric connectivity, $c_{ij} \equiv c_{ji}$, one would consider the energy function

$$E' = -\frac{1}{2} \sum_i^N \sum_{k=1}^S h_i^k \sigma_i^k + U \sum_i^N \sum_{k=1}^S \sigma_i^k. \quad (9)$$

Due to the moving thresholds, however, no energy or free-energy can be defined, to also describe how the thresholds themselves depend on the activity. Nevertheless, we are not so interested in the model in general, but rather specifically in the regime in which neuronal dynamics is faster than threshold dynamics, both intended at the local population level, i.e. when $\tau_3 > \tau_2 \gg \tau_1$ (the *slowly adaptive regime*).

In the slowly adaptive regime the attractors, which were asymptotically stable without adaptation, gradu-

ally become unstable. By going to the limit in which τ_1 is much shorter than the slow τ_2 and τ_3 times, one can conceptualize dynamics as occurring over completely separate time scales.

In a time of order τ_1 the $\{\sigma\}$ configuration relaxes towards an attractive metastable state, determined by nearly constant values for the thresholds, following a continuous time version of the discrete-time updating dynamics mentioned in Sec. II A. Having found a metastable state, an "attractor ruin" in the language of Tsuda [20], the $\{\sigma\}$'s can be rigidly entrapped there, especially if thermal noise is very low, $\beta \gg 1$. At this point neural dynamics becomes almost stationary, while thresholds continue to evolve, albeit slowly. On a time scale τ_2 , and with nearly constant $\{\sigma\}$'s, the thresholds change linearly until they destabilize the $\{\sigma\}$'s again. In this joint limit $\tau_2, \tau_3 \gg \tau_1$ and $\beta \gg 1$, then, and again assuming symmetric connectivity, one can describe separately the two dynamic modes, each governed by a distinct functional.

The rapid transient dynamics towards an attractor minimize the classical Energy

$$\begin{aligned} E(\{\sigma(t)\}|\{\theta\}) = & -\frac{1}{2} \sum_{i,j \neq i}^N \sum_{k,l=1}^S J_{ij}^{kl} \sigma_i^k \sigma_j^l - \frac{w}{2} \sum_i^N \left[\sum_{k=1}^S (\sigma_i^k)^2 - \frac{1}{S} \left(\sum_{k=1}^S \sigma_i^k \right)^2 \right] + \\ & + \sum_i^N \sum_{k=1}^S \left\{ (U + \theta_i^0 + \theta_i^k) \sigma_i^k + \int_0^{\sigma_i^k} [r_i^k(\{\sigma_i^1, \dots, \varrho^k, \dots, \sigma_i^S\}) - U - \theta_i^0] d\varrho^k \right\} \end{aligned} \quad (10)$$

where we have just added the extra local feedback term (the second in the first row), and the thresholds appear as parameters, in the second row, while the second term in the second row follows the formulation proposed for quasi-binary graded response units by Hopfield [44]. This last term can be rewritten, considering that $r_i^k = U + \theta_i^0 + (1/\beta) \ln(\sigma_i^k/\sigma_i^0)$ and carrying out the integrals, as

$$\begin{aligned} E(\{\sigma(t)\}|\{\theta\})|_{\text{graded response term}} = \\ = \frac{1}{\beta} \sum_i^N \sum_{k=1}^S \left\{ \sigma_i^k \ln \frac{\sigma_i^k}{\sigma_i^k + \sigma_i^0} + \sigma_i^0 \ln \frac{\sigma_i^0}{\sigma_i^k + \sigma_i^0} \right\} \end{aligned} \quad (11)$$

which clarifies its nature as an "entropy" term.

The slow adaptation dynamics minimize instead the Adaptation function

$$\begin{aligned} A(\{\theta(t)\}|\{\sigma\}) = & \frac{1}{2} \sum_i^N \sum_{k=1}^S [(\theta_i^k)^2 - 2\theta_i^k \sigma_i^k] + \\ & + \frac{1}{2} \sum_i^N \left[(\theta_i^0)^2 - 2\theta_i^0 \sum_{k=1}^S \sigma_i^k \right], \end{aligned} \quad (12)$$

where now the activation variables appear as parameters.

Note that if we assume that over certain temporal intervals the thresholds $\{\theta\}$ are roughly constant, we have

$$\begin{aligned} \frac{dE}{dt} = & \sum_i^N \sum_{k,l}^S \frac{\delta E}{\delta \sigma_i^k} \frac{\delta \sigma_i^k}{\delta r_i^l} \frac{dr_i^l}{dt} = -\tau_1 \sum_i^N \sum_{k,l}^S \frac{dr_i^k}{dt} \frac{\delta \sigma_i^k}{\delta r_i^l} \frac{dr_i^l}{dt} = \\ = & -\beta \tau_1 \sum_i^N \sum_{k,l}^S \frac{dr_i^k}{dt} \sigma_i^k (\delta_{kl} - \sigma_i^l) \frac{dr_i^l}{dt} \leq 0. \end{aligned} \quad (13)$$

because the matrix $\sigma_i^k (\delta_{kl} - \sigma_i^l)$, which has determinant $\Delta = \prod_{k=0}^S \sigma_i^k$ and trace $\text{Tr} = \sum_{k=1}^S \sigma_i^k (1 - \sigma_i^k)$ can be seen by induction to be positive definite; whereas if we assume that over other temporal intervals the attractor variables $\{\sigma\}$'s are roughly constant, we have instead

$$\begin{aligned} \frac{dA}{dt} = & \sum_i^N \left[\sum_k^S \frac{\delta A}{\delta \theta_i^k} \frac{d\theta_i^k}{dt} + \frac{\delta A}{\delta \theta_i^0} \frac{d\theta_i^0}{dt} \right] = \\ = & -\sum_i^N \left[\tau_2 \sum_k^S \left(\frac{d\theta_i^k}{dt} \right)^2 + \tau_3 \left(\frac{d\theta_i^0}{dt} \right)^2 \right] \leq 0. \end{aligned} \quad (14)$$

Therefore whenever the system is falling or jumping into a new attractor the dynamics are effectively governed by

the energy function, while during the protracted periods of permanence in the attractor, it is the adaptation function to be minimized, while the energy grows again, preparing for the next jump, as illustrated in Fig. 3.

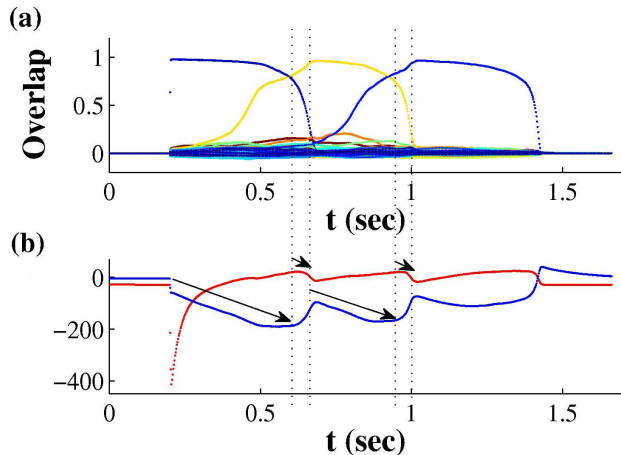


Figure 3: (Color online) (a) Latching transitions; (b) corresponding value of the Energy (red curve) and Adaptation function (blue curve). During a transition, on a fast time scale, the Energy decrease drives the system to a new attractor state. Then, on a slower time scale, Adaptation destabilizes the attractor, pushing the network towards a new transition. As in Fig. 1, time is expressed in *sec*, assuming $\tau_1 = 10$ msec.

A set of simulations were conducted, however, considering what may be called a *rapidly adapting regime*, when in particular τ_3 is very short, $\tau_3 \ll \tau_1 < \tau_2$, which has the computational advantage that rich dynamics unfold within limited CPU time. Approximate analytical considerations derived from those applicable to the slowly adapting regime can be extended to this rapidly adapting regime, which provides a sort of control case, although physiologically less plausible (it would model fast inhibition, but not its important slow component [39]).

D. Correlated patterns

With the introduction of the adaptive thresholds, the network eventually slips away from any attractor it may have retrieved. A free association is said to occur if the system then jumps into a new attractor. As shown in [45], such a jump is facilitated if there is a second memory pattern correlated with the first one. Indeed, while individual units are adapting and slowly changing their activation, those that show the same activation in both patterns act as a cue for the retrieval of the second memory.

To generate a set of correlated patterns we use the algorithm presented in [24]. The algorithm is comprised of a two-step procedure. In the first step a number of

mutually uncorrelated vectors, called *factors*, are established. Each factor can influence the activation of some of the units, by "suggesting" a particular state. In the second step the competition among these factors determines the final activation state of each unit in each of the p patterns.

In particular each factor acts on a distinct, but not exclusive, subset of Na_f units with a strength that varies proportionally to the relevance of the factor itself and to a random number ranging from 0 to 1, which is nonzero with probability a_{pf} . The relative strength (relevance) of the factors decreases exponentially, on average, from the most relevant factor downward. The exponent ζ can be varied parametrically from a small value, producing nearly equivalent factors, to large ones, $\zeta \approx 1$, resulting in the dominant influence of the first few factors. Each pattern has therefore a probability a_{pf} to be influenced by a given factor. After all the relevant factors have expressed their suggestions on an activity direction among the S possible ones, a full activation ($\sigma_i^k = 1$) is assigned, in the state k most activated by the factors, to the Na most activated units (most activated in the selected direction), while the remaining $N(1-a)$ ones are set in the inactive state. In this way, a set of p patterns is produced with exactly $N(1-a)$ units in the inactive state and Na units fully aligned with one of the S active states.

The algorithm thus produces a set of patterns with a certain degree of correlation. The correlation can be tuned acting particularly on two parameters: a_{pf} and the exponent ζ of the strength. Increasing the proportion of factors active in a pattern, a_{pf} , the average correlation between pairs of patterns increases almost linearly, as measured by the averages of the number of units active in the same state in both patterns, N_{as} , or in different states, N_{ad} , or both inactive in the pair, N_{00} , from the reference values taken for randomly generated patterns (not shown). The ζ exponent acts instead on the relative relevance of each factor. The downstream effect on the averages of N_{as} , N_{ad} and N_{00} is non-monotonic, as the average correlation increases with ζ for ζ small, but then it decreases; and its overall magnitude is limited anyway (as shown in Fig.4). The effect is more salient on the variability of the smaller among these numbers, around their averages; in particular the average of N_{as} is quite small (just above 4 units with the parameters of Fig. 4), and its variability from pair to pair is considerably larger than the average. In other words, the algorithm produces patterns that *overall* roughly cover the same portion of their high-dimensional pattern space, with similar average correlations. But when a limited number of factors are relevant *local* relations are altered, with patterns typically having a neighborhood of other patterns with which they share many units active in the same state, and remote patterns with which they have no units in common, active in the same state. It is a non-transitive type of neighborhood (as the active units that define it vary from pair to pair), so it does not imply any clustering in pattern hyper-space.

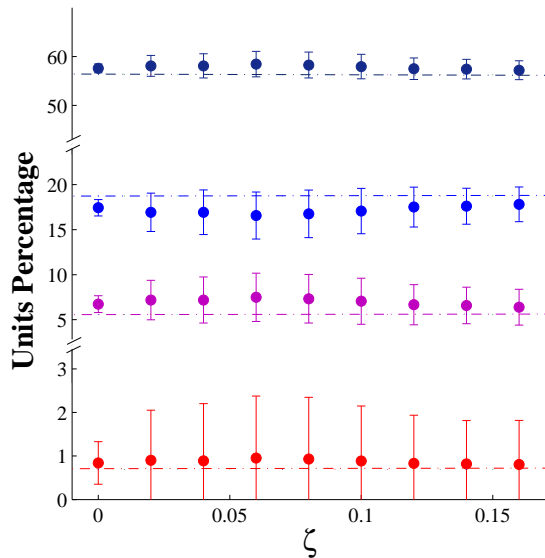


Figure 4: (Color online) The generation algorithm produces a set of correlated patterns, whose correlation can be varied with the ζ exponent. With Potts units, where each can assume a variety of active states, to gauge the correlation between different patterns (network configurations with a fixed number of active units) one has to take in account the number of units shared by the two patterns in the same active (N_{as} , red, bottom) and in the inactive (N_{00} , black, top) state; the number (N_{ad} , pink, third from top) of units active in two different states; and that ($N_{a0} = N_{0a}$, blue, second from top) of units active in a configuration and inactive in the other. The horizontal lines refer to the reference values of N_{as} , N_{ad} , $N_{a0} = N_{0a}$ and N_{00} , expressed as percentages, in the case of randomly generated patterns, with $a_{pf} = 0.3$, $a = 0.25$, $N = 600$, $p = 140$ and $S = 9$. The effect of ζ on the average of these numbers is limited, as shown here in percentages, but intermediate values of ζ greatly amplify the variability of N_{as} around its average.

In [45] we focus especially on the effect of the correlation between two patterns on the dynamics of a single latching transition. Here, instead, we aim to study how structural network parameters influence the dynamics of the whole latching sequence. Therefore, for the sake of simplicity, in most of the analyses we use a set of randomly generated patterns ($a_{pf} = 0$ and $\zeta = 0$), leaving to a brief section the discussion of how correlations alter the latching phase diagram. It is clear, however, that having more "neighbors" to any current pattern, with non-zero a_{pf} and ζ , facilitates latching transitions and also determines which transitions are likely to occur.

For random patterns, their average correlation is set by the sparsity a and the number of active states S . Indeed, two patterns on average share $N(1-a)^2$ inactive units and they have $2Na(1-a)$ units active in just one of the two patterns, Na^2/S units active in the same state and $N(S-1)a^2/S$ units active in a different activation state (the reference values in Fig. 4).

E. Storage capacity

A first observation, about the correlation between patterns, is that increasing it strongly affects the network retrieval performance, i.e. its storage capacity.

The capacity of the Potts network to store information, as with any neural network, is limited. For an autoassociative network, in particular, there is typically a maximal number p_c of patterns, called its *storage capacity*, that can be stored without significantly affecting the ability of the network itself to retrieve each of them from a partial cue [2]. Exceeding this number, at the presentation of the cue the network, once driven solely by internal (recurrent) inputs, reaches a state uncorrelated with any of the memory patterns, or only partially correlated with the memory to be retrieved or, even, with a "wrong" one.

In [26] the storage capacity p_c is analyzed for a variety of Potts autoassociative networks. For each of those, the capacity limit is studied as a function of the global parameter a , quantifying the sparsity of the memory patterns and indirectly of the network activity, and of the local parameter S , the number of Potts states.

The network considered here is a graded-response and adapting variant of what in that analysis is called a "sparse" network (irrespective of activity becoming really sparse only in the limit $a \ll 1$). In the further "thermodynamic" limit $N, p, \beta \rightarrow \infty$, with $\alpha \equiv p/N$ finite, its storage capacity is found to scale, for $a \ll 1$, as

$$p_c \simeq \frac{NS^2}{4a \ln\left(\frac{2S}{a\sqrt{\ln(S/a)}}\right)} \quad (15)$$

for a fully connected network or, for a diluted network as

$$p_c \simeq \frac{CS^2}{4a \ln\left(\frac{2S}{a\sqrt{\ln(S/a)}}\right)}, \quad (16)$$

which are closer estimates than the simple $p_c \approx NS^2/a$ or $p_c \approx CS^2/a$, over a broad range of values of a , as they are for standard autoassociative networks [46].

This analytical result is validated by numerical calculations and by network simulations, but it only holds for the storage of random, uncorrelated patterns. One may now consider how an enhanced correlation among patterns affects the storage capacity.

In a symmetric recurrent network in which one can define a Lyapunov or energy function, the correlation among memory patterns influences directly the energy depth of the network attractors, and hence its storage capacity. In a condition in which the stored patterns are few and uncorrelated, during the retrieval of a memory the overlap of the state of the system with that particular pattern is close to one, and close to zero with the others. In this case it has been shown [47] that in the thermodynamic limit and for $T \rightarrow 0$ the attractor configurations of the network exactly match the stored patterns, and

all have the same depth in energy. With correlated attractors, instead, when a memory is retrieved there is a partial activation of some of the other memories, roughly proportional to their correlation with the retrieved one. The energy minima then have a distribution in energy around the mean value.

Approaching the capacity limit one arrives to a condition where the basins of attraction of different memories overlap and eventually fracture and coalesce, forming new "spurious" stable states [48]. The main effect of correlations is similar, to generate overlaps among the attractors. This makes it more difficult for the network to retrieve the stored patterns.

Fig. 5 shows the difference in p_c for sets of patterns with three different correlation levels, as obtained from computer simulations. From the figure, one sees that the presence of correlations, generated by the multi-factorial algorithm, decreases significantly the capacity of the net, when $a \rightarrow 0$, with respect to the randomly correlated case. One notes that, conversely, the effect of correlations is negligible for non-sparse patterns, $a \geq 0.5$, but in this range the actual capacity is sensibly lower than predicted by the asymptotic formula.

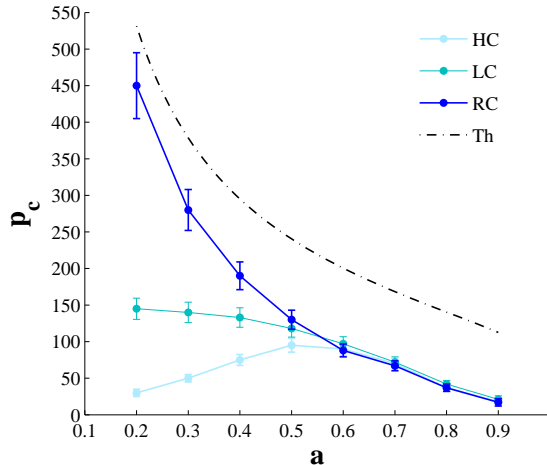


Figure 5: (Color online) Storage capacity of a network of $N = 600$ units with $S = 5$ active states as a function of its sparsity a for three sets of stored patterns: a set of randomly generated patterns (RC: $\zeta = 0.0, a_{pf} = 0.0$, top solid curve), one of weakly correlated (LC: $\zeta = 10^{-6}, a_{pf} = 0.4$, intermediate) and one of highly correlated (HC: $\zeta = 0.1, a_{pf} = 0.4$, bottom curve). The dashed curve shows the theoretical value predicted by Eq. 16, in the $a \ll 1$ limit. In these simulations, $U = 0.5$, $w = 0$, $C = 90$ and $N = 600$.

III. LATCHING SEQUENCES

Once randomly or nonrandomly correlated patterns are stored, one may study the dynamics of the network

elicited by cued retrieval. The initial retrieval is triggered here by a transient signal in the direction of a memory pattern, then the system is left to evolve subject to its own dynamics. What one may observe is shown in Fig. 1. After the first cued retrieval the system adapts and makes a transition to a new attractor. This new retrieved memory adapts in its turn, and acts as a cue for a new transition. Such a behavior is spontaneous and is due just to adaptation and to the correlation among the patterns. In these simulations, transitions are not due to noise since, at the level of interactions among Potts units, the dynamics is deterministic and no noise is included in the equations (T represents instead noise internal to each patch, in the interactions among the neurons subsumed into the effective Potts description). At the same time the sequence of retrieved patterns is not defined *ad hoc*, and it does vary with any small variation of the initial conditions. Finally, one may notice that each of the retrieved patterns can appear multiple times in the retrieval sequence, as the increase in the thresholds is just temporary, and each unit can be reactivated and/or return to its former activity state, in due time.

Exploring the space of parameters, a variety of dynamical behaviors are qualitatively distinguished. For some parameters, that we call the *finite latching region*, the complex dynamics, described above as latching transitions, die spontaneously after a certain number of *steps* and the system relaxes into its inactive configuration, the only truly stable one for the network (Fig. 1(b)). From this region, moving a bit in phase space, i.e. modifying the values of some of the parameters, it is possible to induce an increase or a reduction in the length of the latching sequence, and to approach either of two extreme conditions: a *no latching region*, where the system is able to recall the cued memory but not to express any associative transition after that (Fig. 1(a)), and an *infinite latching region*, where the latching process self-sustains indefinitely in time (Fig. 1(c)).

What makes latching begin, and what makes it eventually stop? We first present a qualitative description of the different regimes, or phases, and then analyze their boundaries, or phase transitions, in the simplified situation in which the stored patterns are uncorrelated or, to be exact, just randomly correlated. Structured correlations are then reintroduced by means of computer simulations.

A. The onset of latching

In [45] we focus on the dynamics of a single latching transition, in order to clarify the internal mechanisms that enable the retrieval of a second pattern and to study qualitatively the variety of transitions arising from these dynamics. The analysis shows that, apart from the somewhat pathological case of oscillations between nearly overlapping attractors, latching transitions may be distinguished between random ones, dubbed *low*

transitions for the low value of the overlap between the latched patterns, and those driven by positive correlations, *high transitions*.

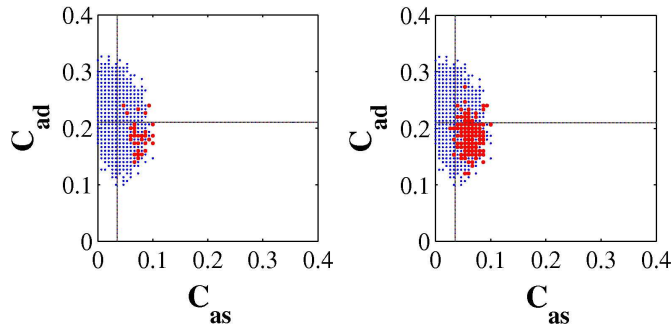


Figure 6: (Color online) Latching transitions (darker, red dots) increase rapidly when the local feedback coefficient increases from $w = 0.35$ (left) to $w = 0.45$ (right). In either case, in the $C_{as} - C_{ad}$ correlation plane, transitions occupy a sub-region of that spanned by correlations among all the stored attractors (lighter, blue dots): they tend to occur only between pairs of patterns with a fraction of active units in the same state above the average value $C_{as} = a/S$ (vertical line) and a fraction in different states below the average value $C_{ad} = a(S-1)/S$ (horizontal line). In these simulations, $a = 0.25$, $S = 7$, and the stored patterns were only randomly correlated.

Following this line, we produce p retrieval sequences by initially cueing each time a different stored attractor. Fig. 6 shows, for those cases in which a transition is produced, the correlation parameters C_{as} and C_{ad} between pairs of latched patterns (red dots) with respect to all the pairs (blue dots). C_{as} is the fraction of units, among the Na active ones in the first pattern, that are also active in the second and in the same state, while C_{ad} is the fraction of the same units that are also active, but in a different state; the complementary fraction $1 - C_{as} - C_{ad}$ are not active in the second pattern of the pair. One can see that, in agreement with [45], the transitions cluster in a region of the correlation plane. With these parameters (in particular, a relatively high threshold U), despite the presence of many patterns sharing a low value of C_{as} , no "low" transitions are observed. The percentage of patterns that give rise to a latching transition clearly increases going deeper into the finite latching region (Fig. 6(b)).

Rather than asking whether transitions occur with a given correlation between the latched patterns, however, we want to focus here on how their occurrence depends on the parameters of the network. In order to study more systematically the border between the no latching and the latching regions, in particular, we vary the number of active states S , the number of stored patterns p , the network connectivity C and the effective temperature T , and extract from simulations the lowest value of the local feedback term w that enables latching transitions (Fig. 7). Note that if these structural parameters are

such that the network straddles the no latching and the finite latching region, transitions may occur or not, for the same parameters and even for the same set of patterns, depending e.g. on the initially cued configuration. This variability generates the error bars in Fig. 7.

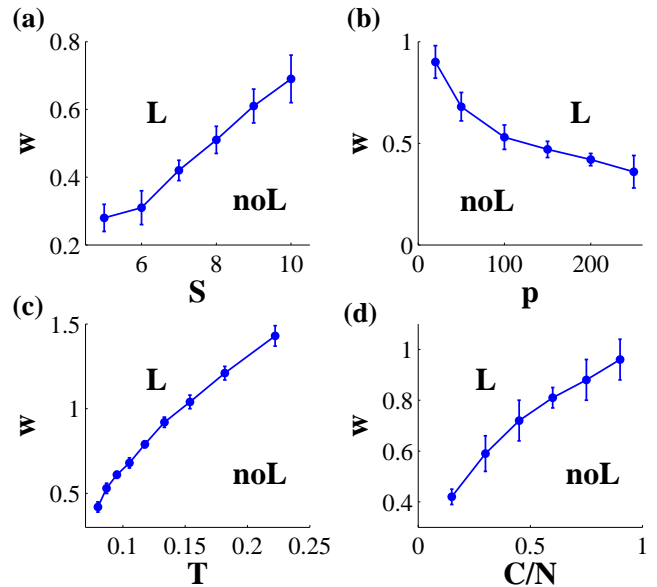


Figure 7: The minimal value of the local feedback coefficient w required to start latching is found with simulations to grow with T (sub-linearly) and with S (linearly or slightly supra-linearly); and to grow approximately with $\sqrt{C/p}$. The storage load (p) thus facilitates latching, whereas local noise, the number of states and the connectivity all obstruct it. When not varied explicitly, the parameters are set at $p = 200$, $S = 7$, $C = 90$, $N = 600$ and $T = 0.08$.

B. Finite latching

To better understand how network parameters affect sequence length, we fix a set of parameters and let the system evolve until it reaches a quiescent condition. If, after $6 \cdot 10^5$ updates of the whole network, equivalent to a time $1.8 \cdot 10^5 \tau_1$, the system is still active, the simulation is terminated.

Fig. 8 shows the average length of the sequences so generated, in relation to the variation of several network parameters: w , S , C , T , τ_2 and τ_3 , while the dependence on p is shown in the next Figure.

Focusing on the parameters defining network architecture, it turns out that an increase in memory load (Fig. 9(a)), a reduction in the number of states of each unit (Fig. 8(a)) and a reduction in the connectivity (Fig. 8(d)) all give rise to an increase in sequence length. Such increase is also produced reducing the effective temperature T (Fig. 8(c)) as well as increasing the local feedback term w (Fig. 8(b)). It is important to note that, for

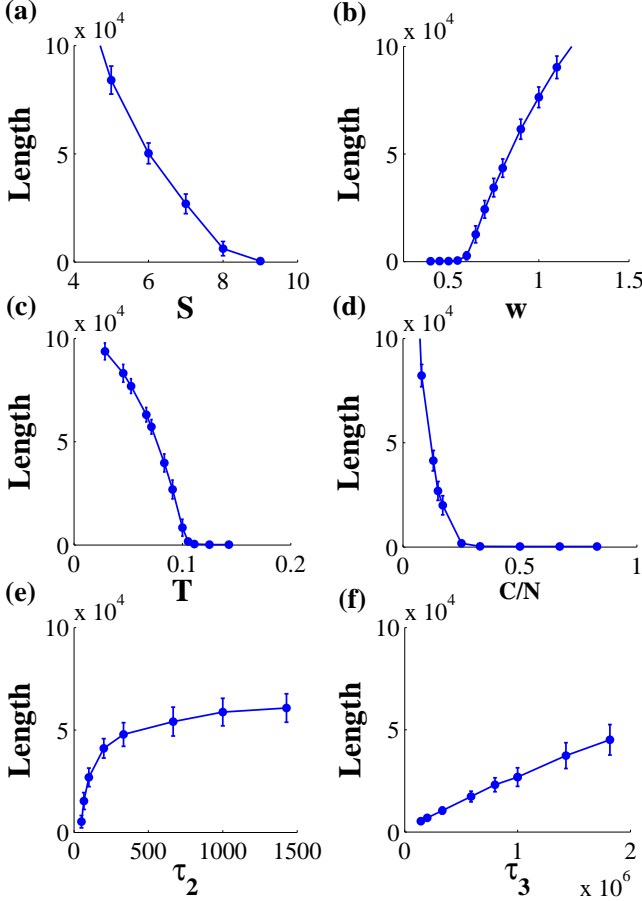


Figure 8: Sequence length (in units of τ_1) as a function of various parameters. The connectivity C , number of attractor states S and local noise T are confirmed to obstruct latching, whereas larger values of the local feedback w (and storage load p , see Fig. 9(a)) prolong it. Moreover, the sequence is found to be stretched almost in direct proportion to the general adaptation time scale τ_3 . When not varied explicitly, the parameters are set at $p = 200$, $S = 7$, $C = 90$, $N = 600$, $T = 0.09$, $w = 0.8$, $\tau_1 = 3.3$, $\tau_2 = 100$ and $\tau_3 = 10^6$.

each of these parameters, there is a critical value beyond which no latching is observed – and often another value, as seen in Fig. 9, beyond which latching appears to proceed indefinitely.

Finally, simulations indicate (Fig. 8(e, f)) that, while the sequence length tends to a plateau by increasing the specific adaptation time scale τ_2 , there is a monotonic, quasi-linear dependence with the generic threshold adaptation time constant, τ_3 .

C. Within the finite latching region

Once the parameter set is fixed, the length of the latching sequence is still not entirely defined. There is indeed

variability due to the initial conditions. As previously mentioned, it is enough to slightly jitter the cued configuration to change completely the sequence of retrieved patterns. While considering the dependence on the memory load p , we focus then on the distribution of lengths seen even for fixed parameters, and due to different initial conditions.

In order to explore the whole configuration landscape, after the storage of the memory patterns, we cue the network in the direction of each of the p patterns, in turn.

In Fig. 9 one can see the distribution of sequence lengths as the network moves from the no latching condition towards the indefinite latching region. Emerging from the no latching phase, as p increases the network begins to latch, initially with only a few jumps between attractor states (the mean sequence length is so short that it is hard to distinguish it from the x -axis). In this condition, the variability in sequence length is spread with a mode close to zero (the observed gap of 200 time steps is approximately the time for the first retrieval to be externally triggered) and a roughly exponential tail (Fig. 9(b, c)). As the average length increases, however, the distribution moves from the exponential to a more normal shape (Fig. 9(d, e)), suggesting a simple relaxation process with a characteristic time scale, which turns out to be just τ_3 (Fig. 8(f)). During this relaxation the system retains a simple memory of the time spent latching, residing in the mean value of the generic adaptation threshold, θ^0 . With time constant τ_3 , in fact, θ^0 approaches its asymptotic value, eventually terminating the latching sequence, when on its way it crosses a certain critical value.

Although so far we have discussed the slowly adaptive regime of Sec. II C, the complexity seen in Fig. 9(a, b, c, d, e) makes it now convenient to review the sequence length distribution in the *rapidly adaptive regime*. As explained in Sec. II B, in such a regime τ_3 is a fast time constant which can be interpreted as characterizing the immediate regulation of the mean activity level locally in the cortex, by fast GABA_A inhibitory mechanism. In such a regime, as prescribed by Eq. 8, the threshold θ_i^0 imposed on each Potts unit tracks the activation level $\sum_{k=1}^S \sigma_i^k$ of the unit itself with speed $1/\tau_3$. Then, by the time the next latching transition occurs, the threshold almost equals the unit activation itself, and it is removed as a distinct variable, thus simplifying the dynamics of the system.

Despite the increase in the sequence lengths, common to both regimes, in the rapid adaptation regime we find a simpler scenario. For a fixed p the length distribution appears quasi-exponential (Fig. 9(g, h, i, l)). This trend is characteristic of an essentially "memory-less" process, in which at each step the probability to fall into the quiescent state, and terminate the sequence, is roughly the same. For higher p , (Fig. 9(i, l)), one notes the appearance of potentially "infinite" sequences, although we terminate them after 3×10^4 steps, thus making the average

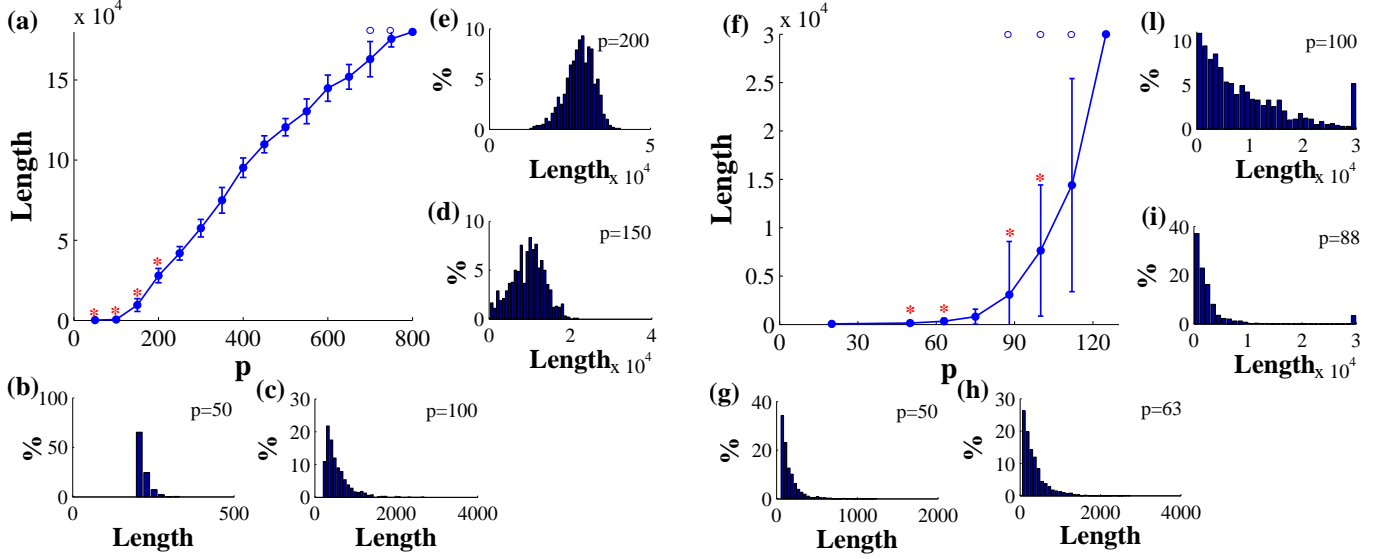


Figure 9: Sequence length, in units of τ_1 , with respect to the number p of stored patterns when the network is in a slow adaptation regime ($\tau_1 = 3.3, \tau_2 = 100, \tau_3 = 10^6$) (a) and in a fast adaptation regime ($\tau_1 = 20, \tau_2 = 200, \tau_3 = 10$) (f). (b, c, d, e, g, h, i, l) Distribution of sequence lengths for a fixed p (at the values indicated by the stars). In the slow adaptation regime, at latching onset, the distribution of lengths is strongly skewed towards low values, with a tail approaching an exponential or Poisson shape (b, c); as p grows and the mean length increases, the distribution is increasingly normal (d, e). The crossover appears to occur when the mean length starts to increase strongly, and quasi-linearly, with p . In the fast adaptation regime, instead, one portion of the distribution remains quasi-exponential (g, h, i, l), but the error bars increase for $p > 80$, due to the inclusion in the distribution of an increasing fraction of trials with an undetermined length, because they were terminated upon reaching the maximum allowed duration of $3 \times 10^4 \tau_1$ (the occurrence of such trials is denoted by empty circles; allowing for longer duration did not increase the quasi-exponential portion appreciably, given p).

length finite. For the same sets of patterns, depending on the initial condition, hence on the landscape visited by the dynamics, finite and infinite (manually terminated) latching sequences co-occur. The increase in the average length for higher p is then due both to an increase in the length of finite sequences and to an increase in the proportion of terminated sequences (not shown).

The quasi-exponential length distribution earlier observed in the slow adaptation regime close to the no-latching region might have been interpreted as resulting from the tail of a normal distribution, when its mode would be below zero, after its negative component is cut off or squashed towards zero. On the other hand, in the fast adaptation regime, the quasi-exponential tail is all that is seen, because when latching does not stop in a few steps, it proceeds indefinitely. But what can terminate it in a few steps, in either regime? The large variability indicates a process, in which the network sometimes falls in a "trap", i.e. in the basin of attraction of one or few "isolated" memory patterns, out of which it cannot latch further.

To test this hypothesis, for a fixed set of patterns we have run extensive simulations, keeping track of the last attractor visited before latching subsides and the network falls in the globally quiescent state.

Fig. 10 shows for each pattern the probability to be

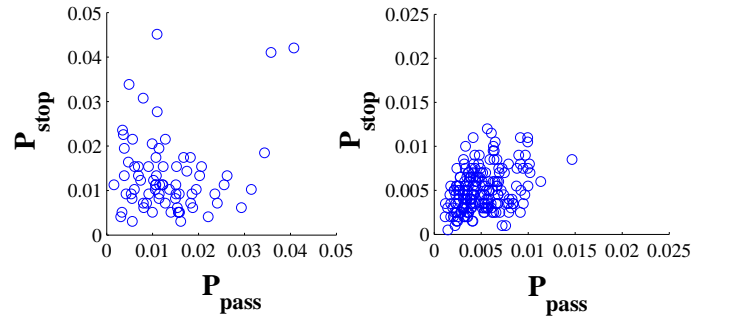


Figure 10: Probability of a pattern to be the last of a latching sequence (P_{stop}) vs. its probability to occur at any position in a sequence (P_{pass}) for the fast (left, $p = 100$) and slow (right, $p = 200$) adaptation regimes.

the last in a latching sequence (P_{stop}), with respect to its probability to occur at any position in a sequence (P_{pass}). In the slow adaptation regime of Fig. 10, right the patterns have a similar probability to occur as the last of a sequence. In contrast, in the fast adaptation regime of Fig. 10, left there seem to be some "preferred" patterns where latching often terminates. Since those patterns are not apparently more likely to be visited than others

before a sequence ends (the relative P_{pass} is not higher than for other patterns) there must be something in their correlational surround that makes them less able to cue the retrieval of other attractors, so that dynamics often stop there. These considerations suggest that the transition to infinite latching occurs when from each basin of attraction a "border crossing" is open to another basin of attraction that the system can latch to, when slipping away of the former.

D. Sustained latching

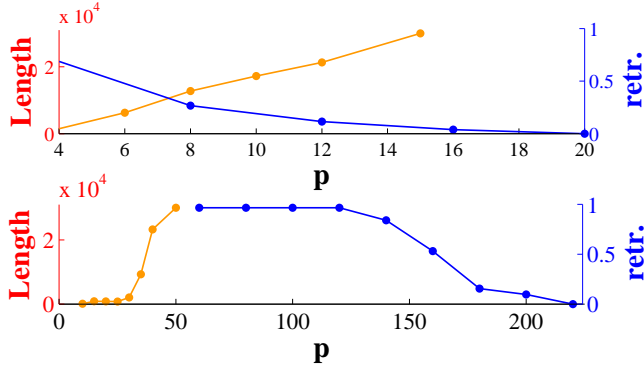


Figure 11: (Color online) Average sequence length (orange ascending curve; left scale, in units of τ_1) and retrieval quality (blue descending curve; right scale) as a function of the storage load, for $S = 3$ (above) and $S = 7$ (below).

As introduced in the previous paragraph, moving further in parameter space, beyond the finite latching region, one seems to enter a domain of infinite sequences, where each latching step is followed by another one, until the simulation has to be terminated. It is important to note that while approaching such infinite latching, the network is also moving towards its storage capacity limit. For some values of the parameters, indeed, the capac-

ity limit, although somewhat ill-defined in a network of moderate size, appears to be crossed before the infinite latching region is reached. If this happens, the system sustains an indefinite noisy activity, without ever stopping and without fully retrieving any pattern. For example, Fig. 11 shows the average sequence length, if not terminated at $3 \times 10^4 \tau_1$, and the fraction of patterns retrieved with an accuracy higher than 90% (a measure of retrieval quality) if network dynamics are initiated by a partial cue. When $S = 7$ one can see that the system reaches the infinite latching region around $p = 50$, while retrievability decreases beyond $p = 120$. Therefore there is a window in the storage load range, in which the system really latches from a retrieved memory to the next, indefinitely. When $S = 3$, instead, the ability to retrieve a cued pattern starts to decline already around $p = 8$, when the length of the latching process is still finite and in fact short. Therefore, for $S = 3$ and with this set of parameters, the network does not ever really latch indefinitely, but rather enters a region of indefinitely protracted noisy dynamics.

E. Rapid adaptation: a percolation type of transition

The observations of Sec. III C suggest that latching dynamics stop when the network runs into an attractor with no "border crossings", or links, to other basins of attraction. Exploring this hypothesis is, however, complicated by the presence of adaptation that continuously modifies the basins and their boundaries.

We consider here the rapidly adapting regime, $\tau_3 \ll \tau_1 < \tau_2$, in which the generic threshold adapts so rapidly that it can be considered as a function quasi-instantaneously set to the value imposed by the $\{\sigma\}$'s, as if due to rapid inhibition. In this case, the effective energy minimized during the rapid transition dynamics does not depend on the θ^0 's any longer, only the θ^k 's for $k \neq 0$ enter as parameters, and the dependence on the σ^k 's is modified by an initial term

$$E = \frac{1}{2} \sum_i^N \left(\sum_{k=1}^S \sigma_i^k \right)^2 - \frac{1}{2} \sum_{i,j \neq i}^N \sum_{k,l=1}^S J_{ij}^{kl} \sigma_i^k \sigma_j^l - \frac{w}{2} \sum_i^N \left[\sum_{k=1}^S (\sigma_i^k)^2 - \frac{1}{S} \left(\sum_{k=1}^S \sigma_i^k \right)^2 \right] + \sum_i^N \sum_{k=1}^S \left\{ (U + \theta_i^k) \sigma_i^k + \frac{1}{\beta} \left(\sigma_i^k \ln \frac{\sigma_i^k}{\sigma_i^k + \sigma_i^0} + \sigma_i^0 \ln \frac{\sigma_i^0}{\sigma_i^k + \sigma_i^0} \right) \right\} \quad (17)$$

necessary to take into account the fast θ^0 dynamics.

To get an impression of the energy landscape visited by the system, without the distortion of adaptation, we consider also E_{land} , that is, Eq. 17 without the contribu-

tion of θ_i^k (while θ_i^0 is kept, though hidden as a function of the $\{\sigma\}$'s).

Fig. 12 shows the overlaps of the system with the stored patterns and the two energy values as a function of time

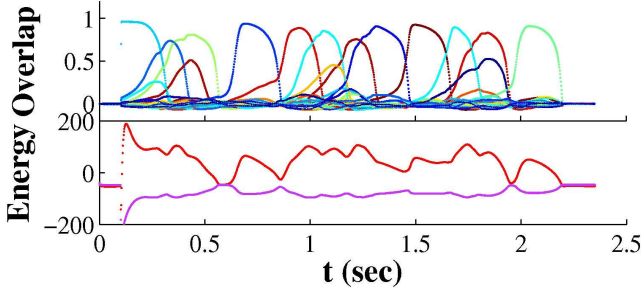


Figure 12: (Color online) Latching transitions and relative value of the total energy (E , higher, red curve of the bottom graph) and of the landscape energy (E_{land} , lower, purple curve). At each attractor transition, the energy decreases while E_{land} has a corresponding peak.

(expressed in *sec*, assuming $\tau_1 = 10msec$). In agreement with Sec. IIB there is a drop in the energy (red line) every time the system has a latching transition, however noisy and far from the ideal of allowing only two non-zero overlaps the transition may be. At the same time, though, E_{land} , the energy related to the landscape (purple line), has a peak. This suggests the presence of an energy barrier that has to be passed each time the system changes basin of attraction. If this hypothesis is correct, we expect the increase in sequence length, approaching the infinite latching region, to be due to a decrease in such barriers. We run then new simulations, crossing the transition to infinite latching by increasing the number of stored patterns.

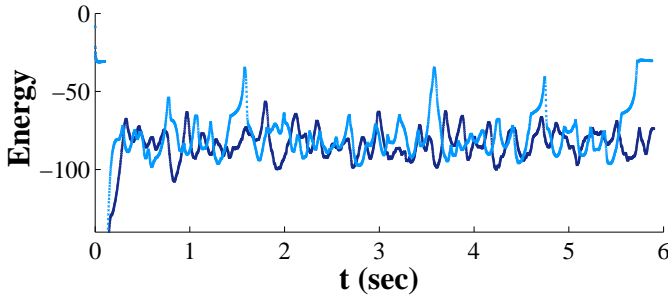


Figure 13: (Color online) Landscape energy profiles (E_{land}) for a network with load $p = 20$ (light blue) and $p = 80$ (darker blue). The increase in the number of stored patterns reduces the highest energy boundaries between attractors.

Fig. 13 shows the energy profile of the system when set below its infinite latching transition (light blue) and above (blue). Comparing the two curves, no major difference is visible between the energy levels of the attractors, i.e. the bottom of the valleys. This appears to exclude a scenario in which the ability to produce an infinite sequence is due to the presence of deeper minima. On the other side, by increasing the number of stored patterns,

the energy boundaries which separate two attractors decrease, especially the highest ones. This confirms our hypothesis and suggests that what actually makes the system stop in a particular attractor is the height of the (landscape) energy boundaries surrounding it.

We can then imagine the energy landscape of the system as a high-dimensional manifold, where the presence of stored patterns is marked by valleys, that correspond to their basins of attraction.

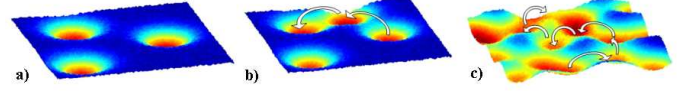


Figure 14: (Color online) 2-dimensional sketches of the energy landscape. (a) When just a few attractors are stored, the system retrieves memories but it is not able to latch. (b) If the number of attractors increases, the basins connect one to the other, allowing some latching transitions. (c) A further increase in the number of stored patterns leads to a condition in which, irrespective of the starting point, the dynamics go on indefinitely in time.

When a few sparse attractors are stored (Fig. 14(a)) the network, if cued towards one of those, works correctly as a memory but without performing any latching transition. The large distance among patterns, indeed, does not allow latching to a new attractor, and once exited from the first attractor the network remains permanently in its stable inactive state (Fig. 1(a)).

If we add new patterns, however, some of their basins of attraction connect (Fig. 14(b)). This lowers the energy boundaries enough, allowing the system to make transitions and creating privileged paths in the landscape. A similar condition is visible in Fig. 1(b), where the system shows latching dynamics, that remain however limited in time. The refractory period due to adaptation, indeed, impedes a new immediate recall of the first memory after the retrieval of the second one. In order to have consecutive transitions, an attractor needs then at least two other correlated attractors, the previous and the following one. The probability of such situation to occur increases with the number of stored attractors, and its inherent variability produce a consequent variation in the length of the latching sequence.

By adding new valleys we arrive finally to a state in which, starting from whatever point in the manifold, the system can always find an infinite path across the landscape (Fig. 14(c) and Fig. 1(c)). This behavior is typical of a *percolation phase transitions* [49, 50]. The increase in the probability of a border crossing, which is higher the more crowded and the wider are the basins corresponding to different memories, and which increases the average sequence length, can then be regarded as the probability of a link in a graph, where the nodes are the attractors. Infinite latching might then require, as in a percolation phase transition, this probability to be above

a certain critical value (leading to the formation of a so called *giant component*).

IV. AN APPROXIMATE ANALYTICAL APPROACH

In order to understand the emergence of distinct phases, let us consider the time derivative of the energy, Eq. 10, when the network is still close to a given attractor ξ^μ and a single unit is changing state, pretending to be in an ideal situation with only clean latching transitions from a single active attractor to the next. We stay now within the slowly adaptive regime. We focus first on a unit i which is currently inactive, i.e. $\xi_i^\mu = 0$ and $\sigma_i^0 \simeq 1$, at the moment just before it starts to increase its activation in a particular direction k , so that a moment later $d\sigma_i^k/dt > 0$. Clearly, the change will lead to a decrease in energy if $dE_i/d\sigma_i^k < 0$, and in that case it is favoured, provided the thresholds can be assumed to stay roughly constant. When still at equilibrium, however, $dE_i/d\sigma_i^k = 0$ (in all S directions), and in fact the small activation values $\sigma_i^k \simeq 0$ of the unit are determined by imposing that all such first derivatives vanish.

Considering that $\sigma_i^0 = 1 - \sum_l \sigma_i^l$, the first derivatives are calculated as

$$\begin{aligned} \frac{dE(\{\sigma_i^m\})}{dt} &= \sum_{k=1}^S \frac{\delta E(\{\sigma_i^m\})}{\delta \sigma_i^k} \frac{d\sigma_i^k}{dt} = \\ &= \sum_{k=1}^S \left[\frac{\delta E_i(\{\sigma_i^m\}, \sigma_i^0)}{\delta \sigma_i^k} - \frac{\delta E_i(\{\sigma_i^m\}, \sigma_i^0)}{\delta \sigma_i^0} \right] \frac{d\sigma_i^k}{dt} \end{aligned} \quad (18)$$

$$\begin{aligned} \frac{\delta E(\{\sigma_i^m\})}{\delta \sigma_i^k} &= - \sum_{j \neq i}^N \sum_{l=1}^S J_{ij}^{kl} \sigma_j^l - w(\sigma_i^k - \frac{1}{S} \sum_{l=1}^S \sigma_i^l) + \\ &+ U + \theta_i^0 + \theta_i^k + T \left(\ln \frac{\sigma_i^k}{\sigma_i^k + \sigma_i^0} - \sum_{l=1}^S \ln \frac{\sigma_i^0}{\sigma_i^l + \sigma_i^0} \right). \end{aligned} \quad (19)$$

The terms after the first, in the partial derivatives, include some (the thresholds) that do not depend on σ_i^k , some (those proportional to w) that are only linear in σ_i^k , and the *entropy* terms, proportional to T , that are highly nonlinear in σ_i^k , going from $-\infty$ to $+\infty$ as σ_i^k goes from 0 to 1. The first term, that reflects the coupling to the other units, $-\sum_{j \neq i}^N \sum_{l=1}^S J_{ij}^{kl} \sigma_j^l$, is constant with respect to the activation values and thresholds of unit i , but it varies a great deal from unit to unit, even among those we are focusing on, inactive in the current configuration and which are activating if a transition occurs. It can be estimated through a signal-to-noise analysis, by considering separately in the couplings J_{ij}^{kl} the contributions due to the storage of the current memory attractor, μ , of the attractor after the transition, ν , and of all others, $\varrho \neq \mu, \nu$. The signal (largely consistent across those units) is due to the first two, and its average is denoted

as $\bar{J}\sigma$, while the variability is due mainly to the storage of the other memories, and to the incomplete connectivity, $c_{ij} \neq 1$, which makes the signal (and noise) vary depending on which are the exact presynaptic units. The standard deviation of this variability is denoted here as $\Delta_{J\sigma}$.

A. The signal and noise from the network

The signal may be estimated as

$$\begin{aligned} -\langle \sum_{j \neq i}^N \sum_{l=1}^S J_{ij}^{kl} \sigma_j^l \rangle &\equiv -\bar{J}\sigma \simeq \frac{a}{S} - [\Gamma_{as} - \Gamma_{ad}] \frac{a^2}{S^2} (S-1) + \\ &- [\Gamma_{as} - \Gamma_{a0}] \frac{a}{S} (1-a) \end{aligned} \quad (20)$$

where Γ_{as}, Γ_{ad} and Γ_{a0} measure the actual correlation between the current and the forthcoming attractor, relative to their expectation values if they were randomly correlated. Specifically, they denote the ratios between the number of units presynaptic to the given unit that are active in the current attractor and are active in the same state (Γ_{as}), in a different state (Γ_{ad}) or inactive (Γ_{a0}) in the forthcoming attractor, and their network averages, which can be estimated as follows (in the case of randomly correlated patterns)

$$\begin{aligned} \Gamma_{as} &= N_{as} / \langle N_{as} \rangle \\ \Gamma_{ad} &= N_{ad} / \langle N_{ad} \rangle \\ \Gamma_{a0} &= N_{a0} / \langle N_{a0} \rangle \end{aligned}$$

$$\begin{aligned} \langle N_{as} \rangle &= Na \langle C_{as} \rangle = Na^2/S \\ \langle N_{ad} \rangle &= Na \langle C_{ad} \rangle = Na^2(S-1)/S \\ \langle N_{a0} \rangle &= Na \langle 1 - C_{as} - C_{ad} \rangle = Na(1-a). \end{aligned} \quad (21)$$

Therefore on average $\Gamma_{as} = \Gamma_{ad} = \Gamma_{a0} = 1$ and the only signal comes from the current attractor (the first term in Eq. 20). As illustrated in Fig. 6, however, latching transitions are favoured among attractors with $\Gamma_{as} > 1$ and $\Gamma_{ad}, \Gamma_{a0} < 1$, since these values bring down the energy derivative. In particular the major contribution is due to the Γ_{as} term (which is the one that most deviates from unity when correlated patterns are generated by the multifactorial algorithm, see Fig. 4). For p random patterns, the one with the largest number of shared active units in the same state as the current attractor can be estimated, from the tail of a Poisson distribution, to have an excess number that roughly satisfies the equation

$$\ln p = \langle N_{as} \rangle \left[\frac{N_{as}}{\langle N_{as} \rangle} \ln \frac{N_{as}}{\langle N_{as} \rangle e} + 1 \right]. \quad (22)$$

For example, if $\Gamma_{as} \simeq 3$ and $\Gamma_{ad} \simeq \Gamma_{a0} \simeq 1$, we have $-\bar{J}\sigma \simeq -a/S$, that is, the excess correlation with the forthcoming attractor is sufficient to switch the sign of the signal from positive (resisting change) to negative (favouring change).

The variability across units is due to several sources, the most important of which are the presence of the other stored patterns, and the incomplete connectivity. Both these factors generate terms that are on average zero across units, but with a variance that can be estimated to be, respectively,

$$\begin{aligned} \langle \sum_j^N \frac{c_{ij}}{(Ca(1 - \frac{a}{S}))^2} \sum_{\varrho=1}^{p-2} \sum_{l=1}^S (\delta_{\xi_i^{\varrho k}} - \frac{a}{S})^2 (\delta_{\xi_j^{\varrho l}} - \frac{a}{S})^2 \delta_{\xi_j^{\mu l}} \rangle &\simeq \\ &\simeq \frac{a(p-2)}{CS^2} \end{aligned} \quad (23)$$

$$\begin{aligned} \left[\langle \sum_j^N \sum_{l=1}^S (c_{ij} \delta_{\xi_j^{\mu l}})^2 \rangle_{c_{ij}} - \sum_j^N \sum_{l=1}^S \langle c_{ij} \delta_{\xi_j^{\mu l}} \rangle_{c_{ij}}^2 \right] \left[\frac{\bar{J}\sigma}{Ca} \right]^2 &\simeq \\ &\simeq \frac{(1 - Ca/N)}{Ca} (\bar{J}\sigma)^2, \end{aligned} \quad (24)$$

which leads to a contribution, taking one standard deviation in the downward direction favouring the transition

$$\begin{aligned} - \left\{ \left\langle \left(\sum_{j \neq i}^N \sum_{l=1}^S J_{ij}^{kl} \sigma_j^l \right)^2 \right\rangle - \left(\left\langle \sum_{j \neq i}^N \sum_{l=1}^S J_{ij}^{kl} \sigma_j^l \right\rangle \right)^2 \right\}^{1/2} &\equiv \\ \equiv -\Delta_{J\sigma} &\simeq \\ \simeq -\sqrt{\frac{a(p-2)}{CS^2} + \frac{a(1 - Ca/N)}{CS^2}} (S\bar{J}\sigma/a)^2 &\simeq \\ \simeq -\frac{\sqrt{a}}{S\sqrt{C}} \sqrt{(p-2) + (S\bar{J}\sigma/a)^2}. \end{aligned} \quad (25)$$

B. The single unit terms

The variability in the coupling term across the S directions also gives the different activation values

$$\sigma_i^k \simeq \exp \beta (\bar{J}\sigma \pm \Delta_J - U - \theta_i^0 - \theta_i^k) \quad (26)$$

which are all small if the unit is largely inactive, with the signal $\bar{J}\sigma$ insufficient to overcome the threshold U . Given the exponential dependence on the fluctuations in the coupling, however, one activation will typically be much larger than the others, which can be taken to hover around their mean value $\bar{\sigma} \simeq 0$. The stability to fluctuations that activate the unit further in the direction in which it is leaning is determined by the stability matrix

$$\begin{aligned} \frac{\delta^2 E(\{\sigma_i^m\})}{\delta \sigma_i^k \delta \sigma_i^l} &= -w(\delta_{kl} - \frac{1}{S}) + T \left[\delta_{kl} \left(\frac{1}{\sigma_i^k} - \frac{1}{\sigma_i^k + \sigma_i^0} \right) + \right. \\ &\quad \left. + \frac{1}{\sigma_i^k + \sigma_i^0} + \frac{1}{\sigma_i^l + \sigma_i^0} - \sum_{m=1}^S \frac{1}{\sigma_i^m + \sigma_i^0} + \frac{S}{\sigma_i^0} \right]. \end{aligned} \quad (27)$$

When the active states are all activated with the same amplitude, the stability matrix has one eigenmode corresponding to the increase of the common amplitude, which

is safely stable, and $S - 1$ degenerate eigenmodes corresponding to increasing amplitudes in each direction. When instead one state is already more active, and the others are close to $\bar{\sigma}$, one of the $S - 1$ eigenmodes becomes unstable earlier, and it is the one proportional to $\delta_{1k} + (\delta_{1k} - 1)\epsilon^k/S$, where we assume an instability in the direction 1 and the small corrections $\epsilon^k \propto \sigma^k$ for $k \neq 1$ express the difference between considering the full stability matrix and simply the second derivative $\delta^2 E_i(\{\sigma_i^m\})/(\delta \sigma_i^1)^2$. The eigenvalue λ satisfies the equations (dropping the unit index i , and for all $k \neq 1$)

$$\begin{aligned} \beta \lambda &= -\beta w(1 - \frac{1}{S}) + \frac{1}{\sigma^1} + \frac{S}{\sigma^0} - \sum_{l \neq 1}^S \frac{1}{\sigma^0 + \sigma^l} + \\ &\quad - \sum_{l \neq 1}^S \frac{\epsilon^l}{S} \left[\beta \frac{w}{S} + \frac{S}{\sigma^0} - \sum_{m \neq 1, l}^S \frac{1}{\sigma^0 + \sigma^m} \right] \end{aligned} \quad (28)$$

$$\begin{aligned} -\beta \lambda \frac{\epsilon^k}{S} &= \beta \frac{w}{S} + \frac{S}{\sigma^0} - \sum_{l \neq 1, k}^S \frac{1}{\sigma^0 + \sigma^l} + \frac{\epsilon^k}{S} (\beta w - \frac{1}{\sigma^k}) + \\ &\quad - \sum_{l \neq 1}^S \frac{\epsilon^l}{S} \left[\beta \frac{w}{S} + \frac{S}{\sigma^0} - \sum_{m \neq l, k}^S \frac{1}{\sigma^0 + \sigma^m} \right] \end{aligned} \quad (29)$$

where the equation after the first can be taken to determine the $\{\epsilon^k\}$'s as a Taylor expansion in the $\{\sigma^k\}$'s, while the first determines the eigenvalue λ to zero'th order in the expansion, as a function of σ^1 , and, combined with the other equations, the following orders.

At the instability, $\lambda = 0$. This determines ϵ^k to be, to first order in σ^k ,

$$\epsilon^k \simeq \sigma^k (\beta w + 2S/\sigma^0) + O(\bar{\sigma}^2) \quad (30)$$

and gives the further condition

$$\begin{aligned} \beta w \left(1 - \frac{1}{S} \right) &\simeq \frac{1}{\sigma^0} + \frac{1}{\sigma^1} + \\ &\quad - \frac{1}{(\sigma^0)^2} \sum_{l \neq 1}^S \sigma^l \left[\left(2 + \frac{\beta w \sigma^0}{S} \right)^2 - 1 \right] + O(\bar{\sigma}^2) \end{aligned} \quad (31)$$

that has to be satisfied together with $\delta E_i/\delta \sigma^1 = 0$ and, in principle, $\delta E_i/\delta \sigma^k = 0$ for $k \neq 1$ to yield σ^1, σ^k and the dependence of e.g. w on all other parameters at the phase transition. In practice, if one can approximate $\sigma^k \simeq 0$ for $k \neq 1$, one needs to solve only Eq. 31 and $\delta E_i/\delta \sigma^1 = 0$, an approximation which can be improved by inserting a small correction term.

C. Transitions

It is now possible to interpret the distinct phases introduced in Sec. III with the analytical tools just established.

The ability of the system to perform or not a transition is tightly linked with the possibility for some of the inactive units to flip in an active state and trigger a cascade, leading to an attractor transition.

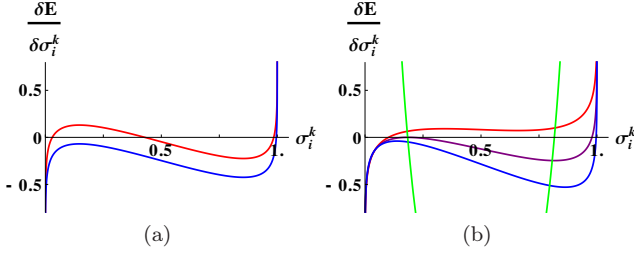


Figure 15: (Color online) Partial derivative of the Energy Functional with respect to σ_i^k . (a) Example of derivative which allows (lower, blue curve) or not (upper, red curve) latching transitions, e.g. because of different values of θ^0 ; (b) derivative in the cases of $w = 1.0$ (upper, red curve), $w = 1.5$ (intermediate, purple curve) and $w = 1.9$ (lower, blue curve), with the other parameters set at $S = 9$, $N = 600$, $p = 140$, $C = 90$, $U = 0.1$, $a = 0.25$ and $\beta = 5$. The light green curve is the stability equation, Eq. 27.

In Fig. 15(a) we plot Eq. 19 with respect to σ_i^k , for two different parameters sets (blue and red curves). In the case of the blue curve the energy derivative is negative until the unit reaches a complete activation. For the red curve, instead, it becomes positive already for small values of σ_i^k . The i unit then, inactive in the retrieved attractor, will activate only if the local maximum of the energy derivative is under the x -axis, as in the blue case. This condition occurs only for some values of the parameters. Eq. 19 depends, indeed, explicitly on w , S , U , T and implicitly, through $\sum_{j \neq i}^N \sum_{l=1}^S J_{ij}^{kl} \sigma_j^l$, on a , C , p , N . The variation of any of them moves and deforms the curve, producing the distinct phases presented in Sec. III.

Among all the parameters, we focus now on the w - T dependence.

D. The w - T phase space

Through computer simulations we explore the w - T phase space. As shown in Fig. 15(b), the qualitative effect of w on the energy derivative is, if increased, to decrease the local maximum and, much more, the minimum of the function.

Fig. 16 shows that, when changing the value of w , the system crosses several possible dynamics scenarios. Increasing w , the system is, respectively, able to retrieve a cued pattern but not to latch to other attractors (*no latching phase*); capable of a finite number of transitions, before getting trapped in the stable silent configuration (*finite latching phase*); capable of an infinite number of transitions (*infinite latching phase*); or trapped in the retrieved configuration (*stable attractor phase*). All this phase transitions correspond to critical conditions in the

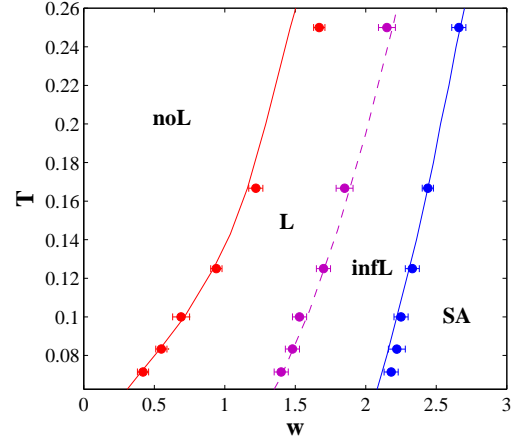


Figure 16: (Color online) The w - T phase space, in the slowly adapting regime. Dots refer to w values, from simulations, that correspond to the beginning of the latching region (red, left), of the infinite latching region (purple, middle) and of the stable attractors region (blue, right). In these simulations $S = 7$, $N = 600$, $p = 140$, $C = 90$, $a = 0.25$ and $U = 0.1$. Solid curves refer to the analytical result (see Sec. IV E), while the dotted curve is just a guide for the eyes. Note that, in this regime, the w -interval between latching onset and the putative percolation transition, to latching sequences of duration proportional to τ_3 , is quite short, within the "error bars" of the simulation data.

energy description.

No Latching - Finite Latching transition. It corresponds to the marginal condition in which the maximum is exactly on the x -axis (Fig. 17(a)). For lower values of w , the derivative is positive and the system cannot latch; for higher values, the derivative is negative and the transition occurs. Whether the maximum is on the axis or not may depend of course on the initial condition, in particular on which pattern is retrieved by the external cue. If it is an early latcher, so to speak, the network may not continue to latch, or it may latch for only a few other steps. If instead w is large enough to allow most attractors to latch, latching will typically last longer. Eventually, all attractors will lead to latching. If this is the case, transitions come one after the other, until the slow θ_i^0 , originally set at 0, slowly increases (as we are focusing now on the slow adaptation regime in which $\tau_3 > \tau_2 \gg \tau_1$), bringing up the whole curve. As soon as the maximum reaches the x -axis, no unit is able to flip any longer, and transitions stop. The w -interval between early and late lachers is quite short, as seen by the standard deviation in the Fig. 16.

Finite - Infinite Latching transition. Since the generic threshold θ_i^0 can take as maximal value $\theta_i^0 = 1$, beyond a certain value of w the growth of θ_i^0 is not sufficient to ever make the maximum positive. For each value of the

temperature T , then, there is a critical value w , beyond which latching transitions never stop.

Infinite latching - Stable Attractors transition. The increase of w , however, has the simultaneous effect of stabilizing network attractors. Therefore, increasing the self-reinforcement term the system reaches a condition in which, after the first retrieval, it remains trapped in the attractor. Analytically we can treat this last case similarly to the previous one, with the only difference that the energy derivative, formally identical to Eq. 19, is now taken with respect to the activation value of an active unit. From Fig. 17(b), it is possible to see that a generic active unit is stable until the Energy derivative has its minimum under the x-axis. Then, when the adaptive threshold θ_i^k increases, following the unit activation, it translates the curve upwards, inactivating the unit as soon as the local minimum crosses 0. If w , however, is too large, the translation due to θ_i^k is not sufficient for the curve to cross the x-axis. There is then a critical value for w , beyond which all network attractors become stable (Fig. 17(a)).

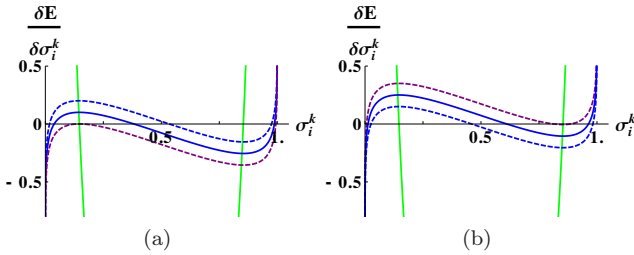


Figure 17: (Color online) Partial derivative of the Energy Functional, mean value (solid curve) \pm standard deviation (dotted curve). (a) Critical condition in which some inactive units (lower, purple curve) are able to flip to an active state and guide a transition; (b) critical condition in which some of the active units (upper, purple curve) get deactivated, thus destabilizing the attractor.

E. Analytical Curves

Having set up an analytical framework, we can now try to explain the boundaries between the different dynamical phases obtained through computer simulations, and shown in Fig. 16.

As described in section IV B, at the instability, Eqs. 19 and 31 should be solved simultaneously, to yield the value of σ_i^k and a relation between the parameters, expressing e.g. w as a function of the others.

Setting all other parameters, and imposing conditions appropriate to the different phases, it is then possible, to solve the system numerically and find the critical values w_c at the phase boundaries. Before analyzing individual phase transitions, we can make a general approximation,

whether unit i is active or inactive and k its future or present activation state: we take $\sigma_i^k + \sigma_i^0 \simeq 1$. The entropy term of Eq. 19 can then be approximated as

$$T \left(\ln \frac{\sigma_i^k}{\sigma_i^k + \sigma_i^0} - \sum_{l=1}^S \ln \frac{\sigma_i^0}{\sigma_i^l + \sigma_i^0} \right) \simeq T \log \frac{\sigma_i^k}{1 - \sigma_i^k}. \quad (32)$$

No Latching - Finite Latching transition. The units that lead this transition are those inactive in the current attractor, that are activated in the next attractor. Among those units, the first to be activated are those that were inactive also before the current attractor, e.g. in the stable condition in which the system is prepared before the external cue that triggers the initial pattern retrieval, so that their θ^k and θ^0 thresholds are still at, or close to, 0. In the unit-to-unit variability expressed by the fluctuations around the mean values in Fig. 17, these units are at the lower end of the range: their activation then reduces the signal, i.e. lowers $\delta E_i / \delta \sigma^1$, for other units, which follow suit. To estimate where is the lower end of the fluctuation range, one may consider the variability expressed by Eq. 25 and the number of units to be activated, which are a subset of the Na units active in the next attractor – those not active already. We find empirically that a good approximation to the $w_c(T)$ curve derived from the simulations is obtained by inserting a coefficient (possibly dependent on N) in front of the s.d. of the signal, and applying a correction to the fraction a of units to be activated, dependent on w and T , so that the mean signal is lowered by a term

$$1.7 \sqrt{\frac{(p-2)(a - \frac{wT}{2})}{CS^2} + \frac{(\bar{J}\sigma)^2}{Ca}} \quad (33)$$

that we leave for future studies to fully justify. The numerical curve thus obtained is in reasonable agreement with that derived from simulations, as shown in Fig. 16. Note that, except for higher T values, the discrepancy between the curve thus obtained and simulation data is within the range of variability of the latter, i.e. within the difference between the very earliest latching, and the putative percolation transition.

Finite - Infinite Latching transition. The boundary between the finite and infinite latching regions is conceptually of the same nature, except that the phase transition occurs when $\delta E_i / \delta \sigma^1$ is lowered so much by the larger w value as to compensate for the thresholds θ^0 and θ^k , which have grown during the latching sequence. Since their final values are difficult to determine, we do not attempt an analytical derivation of the phase boundary.

Infinite latching - Stable Attractors transition. The transition from unstable to stable attractors is related to the inability of any active units to be deactivated by the increase in the adaptive threshold. If a few units go to the inactive state, others follow, with the same chain

reaction process seen above. Therefore the condition for attractors to be stable is that no unit should be able to inactivate. The limit hence correspond to the condition in which, with maximal threshold, and $\theta^k = \theta^0 = 1$, the minimum of Eq. 19 lies on the x-axis. Again, one estimates the height of the $\delta E_i / \delta \sigma^1$ curve for the early deactivating units from the amplitude of the fluctuations in Eq. 25. The result of this approximation is the blue curve in Fig. 16. To produce this curve, we used directly mean θ values resulting from the simulations.

F. Adapting the analysis to the fast adaptation regime

The above analysis can be extended to the situation in which the specific thresholds θ_i^k continue to evolve on a much slower time scale than the activation values σ_i^k , but the generic thresholds θ_i^0 evolve faster, and the limit can be considered in which they adapt almost instantaneously to the activation values. Strictly in that limit, it was shown earlier that one can consider the modified energy functional of Eq. 17, where the θ_i^0 have been integrated out, and there is an additional term $(\sum_k \sigma_i^k)^2 / 2$ for each unit, to account for their rapid adaptation. A pre-factor $\mu < 1$ may also be inserted before the additional term, to take into effective consideration a non-instantaneous adaptation of the generic thresholds, i.e. when $\tau_3 < \tau_1$, but still the two time scales are comparable. It is simpler, though, to restrict ourselves to the limit $\tau_3 \ll \tau_1 \ll \tau_2$.

In this limit, the partial derivative of the energy functional with respect to any of its activation values remains numerically the same as in Eq. 19, with the difference that the simple term θ_i^0 now stands for $\sum_k \sigma_i^k$. As a consequence, the second derivatives that enter the stability matrix change, and they all acquire an additional term equal to 1. The components of the unstable eigenvector in the non-preferred directions are modified into

$$\epsilon^k \simeq \sigma^k (\beta w + \beta S + 2S/\sigma^0) + O(\bar{\sigma}^2); \quad (34)$$

and, importantly, the stability condition now reads, to leading order,

$$\beta w \left(1 - \frac{1}{S}\right) \simeq \beta + \frac{1}{\sigma^0} + \frac{1}{\sigma^1} + O(\bar{\sigma}). \quad (35)$$

The additional β term in Eq. 35 implies that the green curve in Fig. 15 is much higher, hence it takes a higher value of the local feedback coefficient w to reach a phase boundary. The phase diagram of Fig. 16 is basically shifted to the right. Moreover, in the fast adaption regime there is essentially no finite latching region: it is squeezed into a tiny transition strip between the onset of latching, when starting in a favorable initial attractor, and the percolation-type transition to infinite latching, when latching proceeds from any attractor. These analytical expectations are well matched by simulation

results (not shown), although a quantitative agreement, when e.g. $\tau_3 = 0.3 \cdot \tau_1$ does not really vanish, requires inserting a pre-factor to multiply the new term, as indicated above, e.g. $\mu \simeq 0.8$.

G. Latching with correlated patterns

Whereas to treat the fast adaptation regime one has to modify the single unit terms in Eq. 19, to describe the phase diagram appropriate to the storage of correlated attractors one has to consider an enhanced variability, due to correlations, in the local field from the rest of the network, i.e. a larger standard deviation $\Delta_{J\sigma}$. Correlations produce more marked differences between "preferred" and "disfavoured" patterns, in their ability to lead to a latching transition; the distance between the dotted standard deviation curves in Fig. 17 is larger, and consequently latching tends to start off earlier, as a function of an increasing w , and to self-sustain, and eventually to be blocked by attractor stability, later. We do not attempt here an analytical estimate of the $\Delta_{J\sigma}$ produced by our algorithm for generating correlated patterns, but simulations show, in Fig. 18, that the boundaries in phase space are modified as expected. The extent of the latching region is therefore amplified, already in the situation shown, with only weakly correlated patterns. With stronger correlation, the finite latching phase becomes even wider, but to consider the functionality of the associative network one has to remember that retrieval capacity can be seriously impaired, as indicated in Fig. 5.

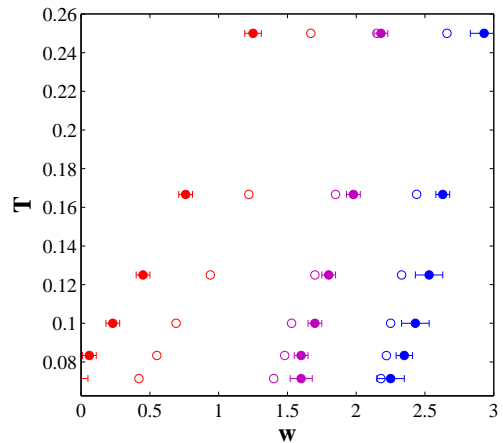


Figure 18: (Color online) The w - T phase space for a correlated set of patterns. Similarly to Fig. 16, dots refer to w_c values dividing the different regions of latching dynamics. Empty dots, the same as in Fig. 16, refer to a randomly generated set of patterns, full dots to a weakly correlated set ($\zeta = 10^{-6}$, $a_{pf} = 0.4$).

V. IMPLICATIONS AND FURTHER WORK

We study here how the introduction of an adaptive element affects the dynamics of an autoassociative neural network. The model draws a sketch of the cortex as a network of Potts units, each of them representing a cortical patch, and interacting with each other through recurrent connections (Sec. II A). The stability of the stored memories, which are encoded in the connections between the units, is weakened by adaptive thresholds, which oppose the permanence in the current attractor. In this way, the network may be pushed out of the basin of attraction of the retrieved memory, and then it can slide towards a new one, performing what we have named a *latching transition* (Sec. II B).

The introduction of adaptation is found to lead to a variety of phases of network dynamics. Depending on its parameters, the network is either unable to perform any transition, remaining inactive after the retrieval of a cued memory; or it does go through some transitions, jumping from memory to memory in a sequential retrieval chain that terminates in the inactive state; or through endless transitions, when the latching process self sustains indefinitely in time (Sec. III). In particular, each of these phases can be reached by changing the number of local attractors assumed to exist in each cortical patch (S), the number of memories stored in the global network (p), the connectivity between cortical patches (C), the local feedback (w) or the noise internal to each patch (T) (Sec. III C).

The infinite latching region is however found not to be always accessible, and its extension in parameter phase space to be restricted by the presence of two other dynamics regions. The increase in memory load of the network, which would favour a sustained latching chain, also impairs the ability of the system to retrieve individual memories. The so-called *storage capacity* (indeed, retrieval capacity) limit can be crossed either after the indefinite latching region is reached, and then it serves as its upper boundary, or before, and then it prevents the system from ever entering such a phase (Sec. III D). On the other hand, attractor stability can also be enhanced by those same manipulations that favour latching. As we saw in section IV D, an increase in the feedback parameter, w , leads to a condition in which attractors are so reinforced, as to become stable. This sets an alternative phase boundary to the latching region.

The introduction of adaptive elements brings a complication in the analytical description. The *effective field* imposed on a unit comes to depend on the time spent by the unit itself in that particular activation state (and in preceding states, as well). As a consequence, it is impossible to write down a function of state governing the dynamics of the system. To bypass this problem, we largely restrict ourselves to a slow adaptation regime and we treat separately the fast attractive from the slow adaptive dynamics (Sec. II B). The time scale separation allow us to focus on the specific instant when attrac-

tors begin to switch, and to derive analytically an approximate phase boundary between the no latching and latching conditions and the indefinite latching and stable attractor conditions, in the w - T phase space (Sec. IV E). The analytical treatment can also be adjusted to fit a type of fast adaptation regime, intended to model rapid local inhibition.

In this somewhat complex scenario, we have finally considered how correlations among patterns may affect network dynamics. With a two-step algorithm, we generate sets of patterns with different distributions of pairwise correlation (Sec. II D). Stronger correlation is found to decrease the storage capacity of the network, and to shrink their basins of attraction (Sec. II E). On the other hand the very same destabilization brought about by correlation makes the system more prone to latching transitions. We sketch the w - T phase space in the case of weakly correlated patterns, and we see that enhanced correlation widens the latching phase, by both lowering the value of w at the no latching-latching boundary and increasing the value needed to stabilize attractors.

A. Applications to analyses of cortical processing

Despite its high level of abstraction, the model can serve as a mathematically well-defined benchmark, where to test general hypotheses about cortical dynamics. An obvious example is any relation hypothesised to exist between the various parameters mentioned above and the duration and characteristics of structured, spontaneous, non-stimulus-driven cortical activity. Experimental paradigms have been used to explore such activity, even in non-human primates, e.g. in the free drawing task developed by Moshe Abeles and coworkers [51]. The behavioural model and the mathematical model can be brought to bear onto each other. Further work is needed, however, to understand how to experimentally manipulate and control any of the parameters, in particular the non-structural ones, including the storage load p , or the local noise level T , or the relative degree of local feedback w , so as to investigate any change induced in the dynamics.

Another issue that can be probed with the model is the relation between the notions of a *global workspace* [52] and of serial *vs.* parallel processing [53]. At the qualitative level of information processing models, the availability of a global workspace can be taken to imply, irrespective of the specialization of individual components [54] that one and only one information processing operation can occur within it at any given time. Hence, a single interconnected network may appear to force serial processing: the lack of segregating boundaries, instead of being a resource, is turned into a nuisance; it imposes a bottleneck. Items that have been processed in parallel along separate sensory pathways can, on their way to distinct motor pathways, get stuck, so to speak, through a router, i.e. a distinct piece of cortical machinery with

ad hoc properties [55], which, like any centralised bureaucrat, acts as a bottleneck [56].

The Potts model, in contrast, with its generic autoassociative architecture, indicates that lack of anatomical segregation is not necessarily equivalent to serial processing. Processing is effectively serial, in the model, when latching dynamics is particularly clean, involving a single global attractor at a time, a situation which does occur in certain regions of parameter space. In other regions, however, dynamics is messier, with several attractors simultaneously active, and latching transitions that involve only part of a structurally complex, but non-segregated network. A rapid transition from attractor A to D may occur while attractors B and C are also active, over units largely distinct from those active in either A or D. Can the system in this case be argued to operate along three separate streams? Hardly so, because the streams are not segregated, and their support basis, i.e. the corresponding subset of active Potts units, changes all the time. The distinction between serial and parallel therefore ceases to be as clear cut as it normally is in engineering systems, and it becomes a *quantitative* rather than qualitative issue. The whole network can be regarded as a single interactive workspace for multifarious parallel processes, like a modern urban society. These aspects require a dedicated study, that adapts information theoretical measures to assess the dependence between successive attractor states [27].

A third domain of applicability of the model is in studying the relation between local and global attractors. In the Potts model, local attractor states are incorporated in the definition itself, which is regarded as a reduced or *effective theory* for a full-fledged neuronal model [25]. The search for experimental evidence for such states has had clearer successes in the hippocampus [5, 9], than in higher visual areas of the cortex [7, 8]. Nevertheless, data compatible with attractor-like behavior is appearing in a variety of preparations, analysed with different approaches, from Hidden Markov Models applied to monkey frontal cortex [57] and to the rat gustatory system [58] to kernel methods applied to multi-unit activity from rat anterior cingulate cortex [59]. A recurrent question is whether such signatures reflect sequences of global attractors states, as have been hypothesized to underly also perceptual decisions in humans, as discrete *moments of thought* [60]. A full mechanistic understanding of such phenomenology seems to require multi-level neuronal network models, where both local and global scales (and possibly intermediate ones) are represented explicitly. The complexity of such model, however, is daunting, so that the Potts model can provide a more tractable approach.

B. Evolutionary perspective

An evolutionary perspective [61] can make use of the Potts model to explore the hypothesis [24] that a criti-

cal step leading to human forms of cognition is a phase transition, incurred as a result of a quantitative change in some of the parameters characterizing the primate (or indeed mammalian) cortex. In this hypothesis, no major structural change differentiates human cortical networks from those of at least fellow great apes, but a qualitatively different functionality of the very same networks may have arisen out of an increase in size, or in connectivity, that has induced a phase transition. The Potts model offers several candidate transitions: from no latching to finite latching, from finite to infinite latching, as well as the percolation-type of transition that, in the slowly adapting regime, follows closely the onset of latching. Identifying the most appropriate scenario is made somewhat complicated by the fact that a plausible cortical model would presumably include both rapidly adapting and slowly adapting thresholds [25], and the two opposite limit cases would have, at the very least, to be combined. This is left for further work. Nevertheless, the notion that spontaneous cortical activity can proceed longer in a larger or more connected brain (not in a scaling relation, but dramatically longer, as after a phase transition) resonates with the notion of infinite recursivity, put forward in linguistics [62] and in other domains [63–65], as we have discussed elsewhere [66].

In this view, the parameters that may constrain cortical functionality are those that are kept at relatively low values, in non-human species, for structural reasons. Indeed, some of the parameters in the Potts model can be thought to reflect quantities that can be freely set at optimal values, and these may include the relative local feedback strength w , the effective noise level T , or the adaptation time scales [67]. Other parameters in the model, including the number of patches N , their connectivity C , the number of attractor states in each patch, S , or the storage load p , reflect cortical quantities that it may be more difficult to act upon, either because increasing their values requires reformulating or extending developmental programmes, or because of their mutual relationships. Thus, for example, increasing the memory storage load is always possible, but it is of little use if, as a result, the system becomes overloaded, beyond its storage capacity, set mainly by the connectivity and by the density of local attractor states. Therefore an increase in p may only be enabled by a concurrent increase in C and/or S [66].

It would be desirable, then, to consider the phase-transition-to-human-cognition hypothesis by studying the phase diagram of the Potts model in relation to the parameters C, N, S, p . Unlike w and T , which affect mainly the single-unit terms, these parameters enter largely (except for a minor feedback role for S) in the signal conveyed by the rest of the network, and in particular they determine its variability $\Delta_{J\sigma}$. The functional dependence of $\Delta_{J\sigma}$ on C, N, S and p is however quite different in different models of the pattern generation process. It can be estimated analytically for randomly correlated patterns, and numerically for patterns generated

by our multi-factorial process, but other models for this key process may be proposed that would lead to rather different effects on the variability of the signal arriving at individual cortical patches, and hence on the phase boundaries of the system. This challenging issue is left for future investigations.

Acknowledgments

We are grateful to Yasser Roudi, Emilio Kropff, Hossein Abbasian and Mohammed F Abdollah-nia, Mo-

ammed Karim Saeed-Ghalati (also for sharing with us their preprint *Optimal region of latching activity in an adaptive Potts model for networks of neurons*) and Andrea Gambassi for several useful discussions and encouragement.

-
- [1] J. Hopfield, PNAS **79**, 2554 (1982).
 - [2] D. J. Amit, H. Gutfreund, and H. Sompolinsky, Annals of Physics **173**, 30 (1987).
 - [3] D. J. Amit, Behavioral Brain Sciences **18**, 617 (1995).
 - [4] E. T. Rolls and A. Treves, *Neural Networks and Brain Function* (Oxford University Press, 1998).
 - [5] T. J. Wills, C. Lever, F. Cacucci, N. Burgess, and J. O'Keefe, Science **308**, 873 (2005).
 - [6] S. Leutgeb, J. K. Leutgeb, C. A. Barnes, E. I. Moser, B. L. McNaughton, and M. B. Moser, Science **309**, 619 (2005).
 - [7] D. J. Amit and S. Fusi, Neural Computation **9**, 1071 (1997).
 - [8] A. Akrami, Y. Liu, A. Treves, and B. Jagadeesh, Cerebral Cortex **19**, 760 (2009).
 - [9] K. Jezek, E. J. Henriksen, A. Treves, E. I. Moser, and M. B. Moser, Nature **478**, 246 (2011).
 - [10] V. Braitenberg, *Cell assemblies in the cerebral cortex* (Springer, 1978).
 - [11] V. Braitenberg and A. Schuz, *Anatomy of the Cortex: Statistics and Geometry* (Springer Verlag, 1991).
 - [12] F. Pulvermüller, Nature Reviews Neuroscience **6**, 576 (2005).
 - [13] T. M. Mitchell, S. V. Shinkareva, A. Carlson, K. M. Chang, V. L. Malave, R. A. Mason, and M. A. Just, Science **320**, 1191 (2008).
 - [14] R. Epstein, Consciousness and Cognition **9**, 550 (2000).
 - [15] P. W. Burgess and T. Shallice, Memory **4**, 359 (1996).
 - [16] T. Shmiel, R. Drori, O. Shmiel, Y. Ben-Shaul, Z. Nadasdy, M. Shemesh, M. Teicher, and M. Abeles, JN Physiol **96**, 2645 (2006).
 - [17] R. Sosnik, M. Shemesh, and M. Abeles, Cognitive Neurodynamics **1**, 341 (2007).
 - [18] M. Abeles, Science **304**, 523 (2004).
 - [19] K. Kaneko and I. Tsuda, Chaos **13**, 926 (2003).
 - [20] I. Tsuda, Neural Networks **5**, 313 (1992).
 - [21] D. O'Kane and A. Treves, Journal of Physics A: Mathematical General **A 25**, 5055 (1992).
 - [22] I. Kanter, Physical Review A **37**, 2739 (1988).
 - [23] D. Bolle, R. Cools, P. Dupont, and J. Huyghebaert, Journal of Physics A: Mathematical General **26**, 549 (1993).
 - [24] A. Treves, Cognitive Neuropsychology **6**, 276 (2005).
 - [25] A. Akrami, E. Russo, and A. Treves, Brain Research **in press** (2011).
 - [26] E. Kropff and A. Treves, Journal of Statistical Mechanics: Theory and Experiment **2005**, P08010 (2005).
 - [27] E. Russo, S. Pirmoradian, and A. Treves, in *Advances in Cognitive Neurodynamics (II)* (Springer, 2011), pp. 111–115.
 - [28] M. E. Costa, F. Bonomo, and M. Sigman, Frontiers in Integrative Neuroscience **3**, 19 (2009).
 - [29] S. Lacey, N. Tal, A. Amedi, and K. Sathian, Brain Topograph **21**, 269 (2009).
 - [30] H. Sompolinsky and I. Kanter, Physical Review Letters **57**, 2861 (1986).
 - [31] D. Horn and M. Usher, Physical Review A **40**, 1036 (1989).
 - [32] M. Herrmann, E. Ruppig, and M. Usher, Biological Cybernetics **68**, 453 (1993).
 - [33] M. Rabinovich, A. Volkovskii, P. Lecanda, R. Huerta, H. D. I. Abarbanel, and G. Laurent, Physical Review Letters **87**, 068102 (2001).
 - [34] R. Metzler and W. Kinzel, Physical Review E **63**, 056126 (2001).
 - [35] M. Timme, F. Wolf, and T. Geisel, Phys. Rev. Lett. **89**, 154105 (2002).
 - [36] D. R. Paula, A. D. Araújo, J. S. Andrade, H. J. Herrmann, and J. A. C. Gallas, Phys. Rev. E **74**, 017102 (2006).
 - [37] C. Gros, New Journal of Physics **9** (2007).
 - [38] G. Tamás, A. Lorincz, A. Simon, and J. Szabadics, Science **299**, 1902 (2003).
 - [39] M. M. Kohl and O. Paulsen, Advances in Pharmacology **58**, 205 (2010).
 - [40] E. Fino and R. Yuste, Neuron **69**, 1188 (2011).
 - [41] M. V. Tsodyks and H. Markram, PNAS **94**, 719 (1997).
 - [42] M. Abdollah-nia, M. Saeedghalati, and A. Abbassian, JS-TAT **in press** (2012).
 - [43] E. Russo and A. Treves, in *Advances in Cognitive Neurodynamics (III)* (in press).
 - [44] J. J. Hopfield, PNAS **81**, 3088 (1984).
 - [45] E. Russo, V. M. K. Namboodiri, A. Treves, and E. Kropff, New Journal of Physics **10**, 015008 (2008).
 - [46] A. Treves and E. T. Rolls, Network **2**, 371 (1991).
 - [47] D. J. Amit, H. Gutfreund, and H. Sompolinsky, Physical Review A **32**, 1007 (1985).
 - [48] Y. Roudi and A. Treves, Physical Review E **67**, 041906 (2003).
 - [49] D. Stauffer and A. Aharony, *Introduction to percolation theory* (Taylor & Francis, 1985).
 - [50] M. K. Phani and D. Dhar, Journal of Physics A: Mathematical and General **17**, L645 (1984).

- [51] T. Shmiel, R. Drori, O. Shmiel, Y. Ben-Shaul, Z. Nadasdy, M. Shemesh, M. Teicher, and M. Abeles, PNAS **102**, 18655 (2005).
- [52] S. Dehaene, M. Kerszberg, and J.-P. Changeux, PNAS **95**, 14529 (1998).
- [53] A. Del Cul, S. Baillet, and S. Dehaene, PLoS Biology **5**, e260 (2007).
- [54] F. Kouneiher, S. Charron, and E. Koechlin, Nature Neuroscience **12**, 939 (2009).
- [55] A. Zylberberg, D. F. Slezak, P. R. Roelfsema, S. Dehaene, and M. Sigman, PLoS computational biology **6**, e100076 (2010).
- [56] C. Pallier, A. D. Devauchelle, and S. Dehaene, PNAS **108**, 2522 (2011).
- [57] M. Abeles, H. Bergman, I. Gat, I. Meilijson, E. Seidemann, N. Tishby, and E. Vaadia, PNAS **92**, 8616 (1995).
- [58] L. M. Jones, A. Fontanini, B. F. Sadacca, P. Miller, and D. B. Katz, PNAS **104**, 18772 (2007).
- [59] E. Balaguer-Ballester, C. C. Lapish, J. K. Seamans, and D. Durstewitz, PLoS Computational Biology **7**, e1002057 (2011).
- [60] M. Graziano, P. Polosecki, D. E. Shalom, and M. Sigman, Frontiers in Integrative Neuroscience **5**, Epub (2011).
- [61] L. Krubitzer and J. Kaas, Current Opinion in Neurobiology **15**, 444 (2005).
- [62] M. Hauser, N. Chomsky, and W. Fitch, Science **298**, 1569 (2002).
- [63] R. Jackendoff, Music Perception **26**, 195 (2009).
- [64] S. J. Hespos and E. S. Spelke, Nature **430**, 453 (2004).
- [65] D. Amati and T. Shallice, Cognition **103**, 358 (2007).
- [66] E. Russo and A. Treves, Biolinguistics **5**, 133 (2011).
- [67] S. Marom, Progress in Neurobiology **90**, 16 (2010).